US009449694B2

(12) **United States Patent**
Paudel et al.

(10) **Patent No.: US 9,449,694 B2**
(45) **Date of Patent: Sep. 20, 2016**

(54) **NON-VOLATILE MEMORY WITH MULTI-WORD LINE SELECT FOR DEFECT DETECTION OPERATIONS**

(71) Applicant: **SanDisk Technologies Inc.**, Plano, TX (US)

(72) Inventors: **Rajan Paudel**, San Jose, CA (US); **Jagdish Sabde**, Fremont, CA (US); **Sagar Magia**, Milpitas, CA (US); **Khanh Nguyen**, Fremont, CA (US)

(73) Assignee: **SanDisk Technologies LLC**, Plano, TX (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **14/477,339**

(22) Filed: **Sep. 4, 2014**

(65) **Prior Publication Data**

US 2016/0071594 A1 Mar. 10, 2016

(51) **Int. Cl.**

G11C 16/10	(2006.01)
G11C 16/08	(2006.01)
G11C 8/14	(2006.01)
G11C 11/56	(2006.01)
G11C 29/26	(2006.01)

(Continued)

(52) **U.S. Cl.**

CPC ***G11C 16/10*** (2013.01); ***G11C 8/14*** (2013.01); ***G11C 11/5628*** (2013.01); ***G11C 29/26*** (2013.01); *G11C 16/0483* (2013.01); *G11C 16/08* (2013.01); *G11C 2029/1202* (2013.01); *G11C 2029/2602* (2013.01)

(58) **Field of Classification Search**

CPC ... G11C 16/10; G11C 16/0483; G11C 16/08; G11C 8/14; G11C 11/5628; G11C 29/26; G11C 2029/1202; G11C 2029/2602

USPC 365/185.11, 185.01, 185.17, 185.13, 365/185.14, 185.12, 185.09

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,070,032 A 12/1991 Yuan et al.
5,095,344 A 3/1992 Harari

(Continued)

FOREIGN PATENT DOCUMENTS

CN	101377960 A	3/2009
EP	2 261 806 A1	12/2010
WO	WO 2007/016167	2/2007

OTHER PUBLICATIONS

Eitan et al., "NROM: A Novel Localized Trapping, 2-Bit Nonvolatile Memory Cell," IEEE Electron Device Letters, vol. 21, No. 11, Nov. 2000, pp. 543-545.

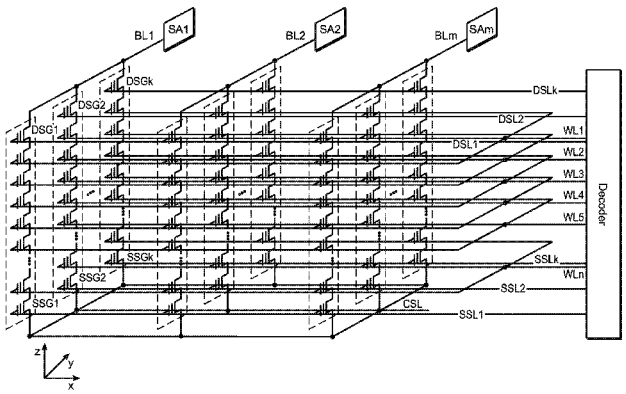
(Continued)

Primary Examiner — Vu Le
Assistant Examiner — Sung Cho
(74) *Attorney, Agent, or Firm* — Davis Wright Tremaine LLP

(57) **ABSTRACT**

A stress mode for use in testing non-volatile memory arrays for a number of types of defects is described. More specifically, a multi-word line select option for a given block can be used for a group of selected word lines to be set to the a programming or other high voltage, while the unselected word lines of the block are set to a pass voltage to minimize electric field differences in order to avoid disturb. For example, a group of selected word lines could number 4, 8 or 16. The multi-word line option can be applied to one block per plane, so that if there are two memory planes, for example, two such blocks can be selected simultaneously for the multi-word line option for those blocks.

15 Claims, 42 Drawing Sheets



(51) **Int. Cl.**

G11C 16/04	(2006.01)
G11C 29/12	(2006.01)

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,335,198	A	8/1994	Van Buskirk et al.	
5,570,315	A	10/1996	Tanaka et al.	
5,602,789	A	2/1997	Endoh et al.	
5,627,784	A	5/1997	Roohparvar	
5,661,053	A	8/1997	Yuan	
5,673,222	A	9/1997	Fukumoto et al.	
5,768,192	A	6/1998	Eitan	
5,822,256	A	10/1998	Bauer et al.	
5,903,495	A	5/1999	Takeuchi et al.	
6,002,612	A	12/1999	Noda et al.	
6,011,725	A	1/2000	Eitan	
6,046,935	A	4/2000	Takeuchi et al.	
6,097,666	A *	8/2000	Sakui	G11C 8/12 365/185.11
6,185,709	B1	2/2001	Dreibelbis et al.	
6,219,286	B1	4/2001	Fuchigami et al.	
6,222,762	B1	4/2001	Guterman et al.	
6,285,597	B2	9/2001	Kawahara et al.	
6,370,075	B1	4/2002	Haeberli et al.	
6,469,934	B2 *	10/2002	de Sandre	G11C 16/3468 365/185.13
6,556,465	B2	4/2003	Haeberli et al.	
6,560,143	B2	5/2003	Conley et al.	
6,760,262	B2	7/2004	Haeberli et al.	
6,922,096	B2	7/2005	Cernea	
7,012,835	B2	3/2006	Gonzalez et al.	
7,030,683	B2	4/2006	Pan et al.	
7,135,910	B2	11/2006	Cernea	
7,158,421	B2	1/2007	Cernea et al.	
7,206,230	B2	4/2007	Li et al.	
7,243,275	B2	7/2007	Gongwer et al.	
7,301,812	B2	11/2007	Guterman et al.	
7,304,893	B1	12/2007	Hemink	
7,345,928	B2	3/2008	Li	
7,366,022	B2	4/2008	Li et al.	
7,368,979	B2	5/2008	Govindu et al.	
7,440,319	B2	10/2008	Li et al.	
7,460,401	B2 *	12/2008	Naura	G11C 16/344 365/185.11
7,554,311	B2	6/2009	Pan	
7,616,484	B2	11/2009	Auclair et al.	
7,616,499	B2	11/2009	Wan et al.	
7,683,700	B2	3/2010	Huynh et al.	
7,716,538	B2	5/2010	Gonzalez et al.	
7,746,707	B2 *	6/2010	Tanaka	G11C 11/5621 365/185.17
7,782,681	B2 *	8/2010	Kim	G11C 16/0483 365/185.22
7,795,952	B2	9/2010	Lui et al.	
7,864,588	B2	1/2011	Betser et al.	
7,969,235	B2	6/2011	Pan	
7,973,592	B2	7/2011	Pan	
8,027,195	B2	9/2011	Li et al.	
8,040,744	B2	10/2011	Gorobets et al.	
8,054,680	B2	11/2011	Matsuzaki et al.	
8,102,705	B2	1/2012	Liu et al.	
8,416,624	B2	4/2013	Lei et al.	
8,726,104	B2	5/2014	Sharon	
8,750,042	B2	6/2014	Sharon	
2001/0004326	A1	6/2001	Terasaki	
2002/0007386	A1	1/2002	Martin et al.	
2002/0024861	A1	2/2002	Chevallier et al.	
2003/0117851	A1	6/2003	Lee et al.	
2003/0147278	A1 *	8/2003	Tanaka et al.	365/185.03
2003/0217323	A1	11/2003	Guterman et al.	
2004/0008553	A1 *	1/2004	Pekny	G11C 29/26 365/201
2004/0022092	A1	2/2004	Dvir et al.	
2005/0024939	A1	2/2005	Chen et al.	
2005/0219896	A1	10/2005	Chen et al.	
2006/0090112	A1	4/2006	Cochran et al.	
2006/0104104	A1 *	5/2006	Park	G11C 16/3468 365/100
2006/0221714	A1	10/2006	Li et al.	
2006/0227602	A1	10/2006	Honma et al.	
2006/0239111	A1	10/2006	Shingo	
2007/0016167	A1	2/2007	Jun et al.	
2007/0030732	A1	2/2007	Micheloni et al.	
2007/0047327	A1	3/2007	Goda et al.	
2007/0171719	A1	7/2007	Hemink	
2007/0230261	A1 *	10/2007	Haufe	G11C 29/50 365/201
2007/0280000	A1	12/2007	Fujiu et al.	
2007/0296391	A1	12/2007	Bertin et al.	
2008/0062765	A1	3/2008	Tu et al.	
2008/0068888	A1 *	3/2008	Kim	G11C 16/0483 365/185.11
2008/0307342	A1	12/2008	Furches et al.	
2009/0058506	A1	3/2009	Nandi et al.	
2009/0058507	A1	3/2009	Nandi et al.	
2009/0063918	A1 *	3/2009	Chen	G11C 8/08 714/721
2009/0100290	A1	4/2009	Nakanishi et al.	
2009/0153230	A1	6/2009	Pan et al.	
2009/0153232	A1	6/2009	Fort et al.	
2009/0190403	A1 *	7/2009	Kwak	G11C 11/5635 365/185.11
2009/0225607	A1	9/2009	Chen et al.	
2009/0295434	A1	12/2009	Umeda et al.	
2009/0315616	A1	12/2009	Nguyen et al.	
2009/0316483	A1	12/2009	Kim et al.	
2009/0323417	A1	12/2009	Takada	
2010/0054019	A1 *	3/2010	Toda	G11C 8/12 365/148
2010/0070209	A1	3/2010	Sai	
2010/0082881	A1	4/2010	Klein	
2010/0091568	A1	4/2010	Li et al.	
2010/0091573	A1	4/2010	Li et al.	
2010/0091578	A1 *	4/2010	Kim	G11C 16/16 365/185.33
2010/0309719	A1	12/2010	Li et al.	
2011/0066793	A1	3/2011	Burd	
2011/0096601	A1	4/2011	Gavens et al.	
2011/0145633	A1	6/2011	Dickens et al.	
2011/0148509	A1	6/2011	Pan	
2011/0182121	A1	7/2011	Dutta et al.	
2011/0194357	A1	8/2011	Han	
2011/0286279	A1 *	11/2011	Lei	G11C 16/3404 365/185.19
2011/0286280	A1	11/2011	Kellam et al.	
2011/0310673	A1 *	12/2011	Cho	G11C 16/0483 365/185.22
2012/0008384	A1	1/2012	Li et al.	
2012/0008405	A1	1/2012	Shah et al.	
2012/0008410	A1 *	1/2012	Huynh	G11C 29/02 365/185.21
2012/0220088	A1	8/2012	Alsmeier	
2012/0224409	A1 *	9/2012	Yan	G11C 13/0007 365/148
2012/0281479	A1	11/2012	Kochar et al.	
2012/0311407	A1	12/2012	Lee et al.	
2013/0028021	A1 *	1/2013	Sharon	G11C 11/5642 365/185.17
2013/0031429	A1	1/2013	Sharon et al.	
2013/0031430	A1	1/2013	Sharon	
2013/0031431	A1 *	1/2013	Sharon	G06F 11/1072 714/719
2013/0107628	A1	5/2013	Dong et al.	
2013/0114342	A1	5/2013	Sakai et al.	
2013/0215678	A1 *	8/2013	Yang	G06F 11/1048 365/185.03
2013/0229868	A1	9/2013	Koh et al.	
2014/0003157	A1	1/2014	Mui et al.	

(56) **References Cited**

U.S. PATENT DOCUMENTS

2014/0136764	A1*	5/2014	Li	G11C 16/08 711/103
2014/0169095	A1	6/2014	Avila et al.	

OTHER PUBLICATIONS

Pan, "Charge Pump Circuit Design," McGraw-Hill, 2006, 26 pages.
Pylarinos et al., "Charge Pumps: An Overview," Department of Electrical and Computer Engineering University of Toronto, www.eecg.toronto.edu/~kphang/ece1371/chargepumps.pdf., 7 pages.
U.S. Appl. No. 13/332,780 entitled "Simultaneous Sensing of Multiple Wordlines and Detection of NAND Failures," filed Dec. 21, 2011, 121 pages.
U.S. Appl. No. 13/101,765 entitled "Detection of Broken Word-Lines in Memory Arrays," filed May 5, 2011, 63 pages.
U.S. Appl. No. 13/193,083 entitled "Non-Volatile Memory and Method with Accelerated Post-Write Read Using Combined Verification of Multiple Pages," filed Jul. 28, 2011, 100 pages.
U.S. Appl. No. 13/280,217 entitled "Post-Write Read in Non-Volatile Memories Using Comparison of Data as Written in Binary and Multi-State Formats," filed Oct. 24, 2011, 110 pages.
U.S. Appl. No. 13/193,148 entitled "Data Recovery for Detection Word Lines During Programming of Non-Volatile Memory Arrays," filed Jul. 28, 2011, 48 pages.
U.S. Appl. No. 11/303,387 entitled "Charge Pump Regulation Control for Improved Power Efficiency," filed Dec. 16, 2011, 25 pages.
First Office Action issued for Chinese Patent Application No. 2011800340192 mailed on Dec. 2, 2014, 7 pages.
First Office Action issued for Chinese Patent Application No. 2011800388647 mailed on Feb. 3, 2015, 12 pages.

* cited by examiner

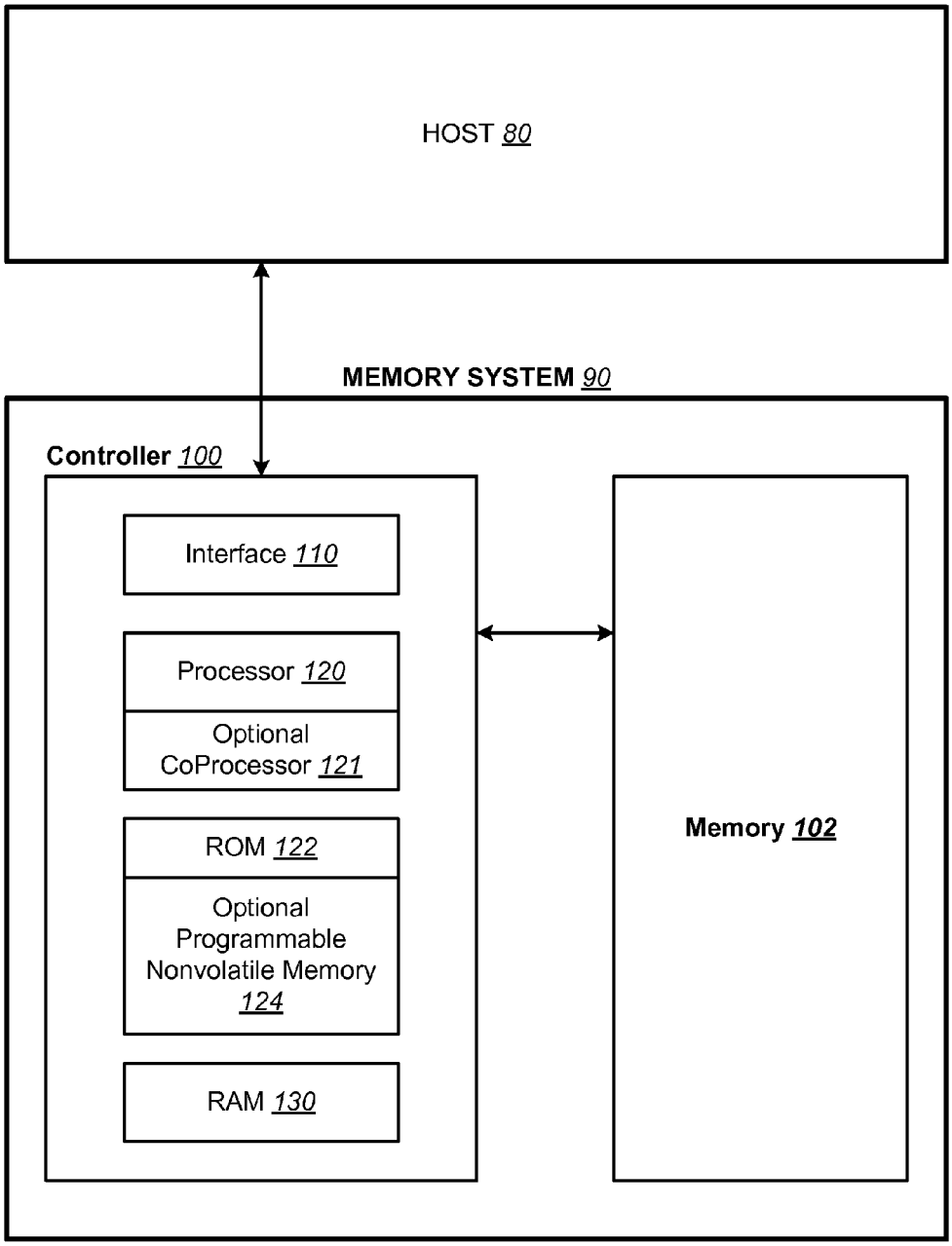

HOST 80
MEMORY SYSTEM 90
Controller 100
Interface 110
Processor 120
Optional CoProcessor 121
ROM 122
Optional Programmable Nonvolatile Memory 124
RAM 130
Memory 102

FIG. 1

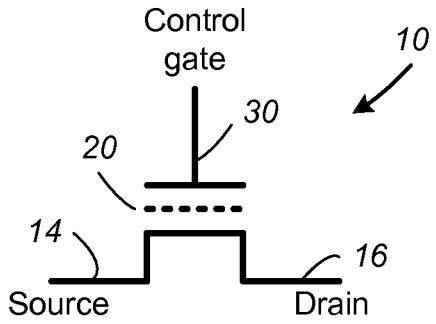

Control
gate
10
30
20
14
16
Source
Drain

FIG. 2

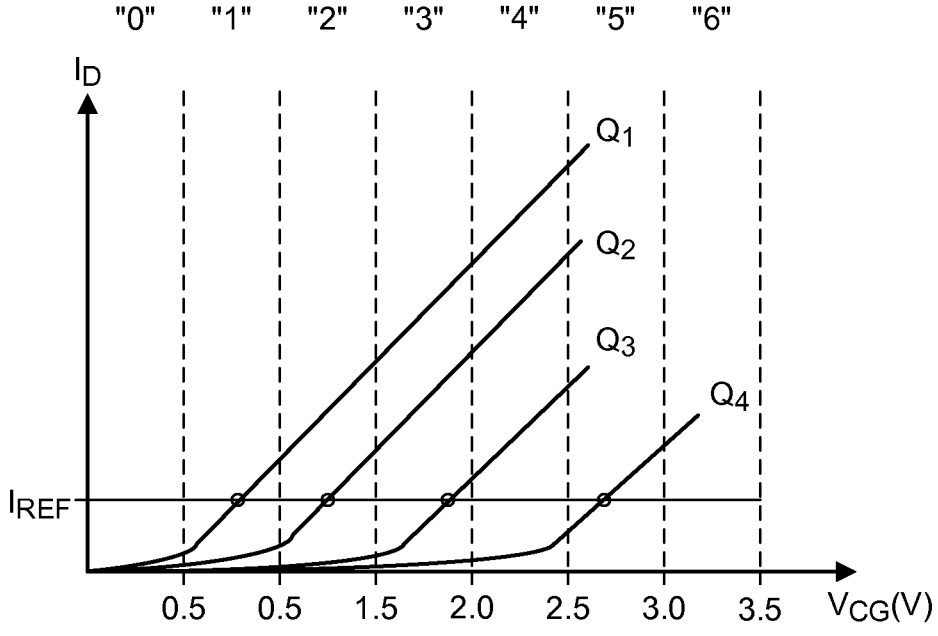

"0"
"1"
"2"
"3"
"4"
"5"
"6"
I_D
Q_1
Q_2
Q_3
Q_4
I_{REF}
0.5
0.5
1.5
2.0
2.5
3.0
3.5
V_{CG}(V)

FIG. 3

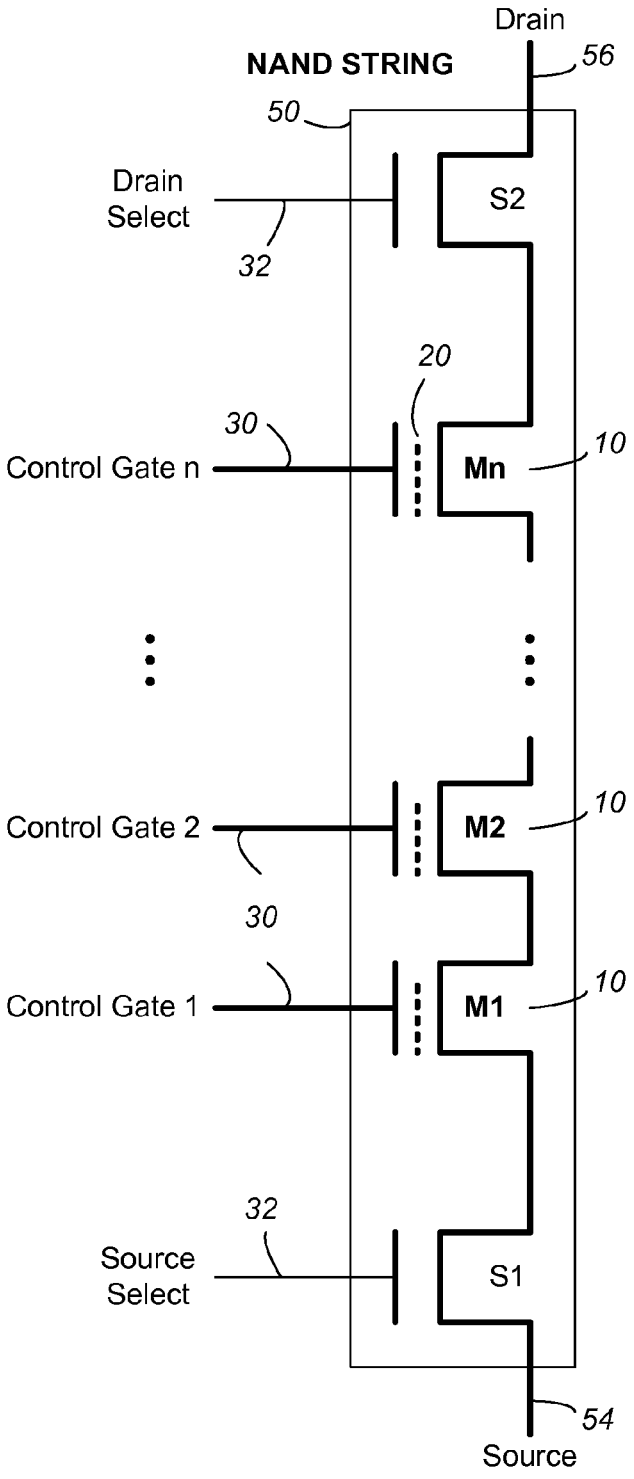

Drain
NAND STRING
56
50
Drain
Select
32
S2
20
30
Control Gate n
Mn
10
Control Gate 2
M2
10
30
Control Gate 1
M1
10
32
Source
Select
S1
54
Source

FIG. 4

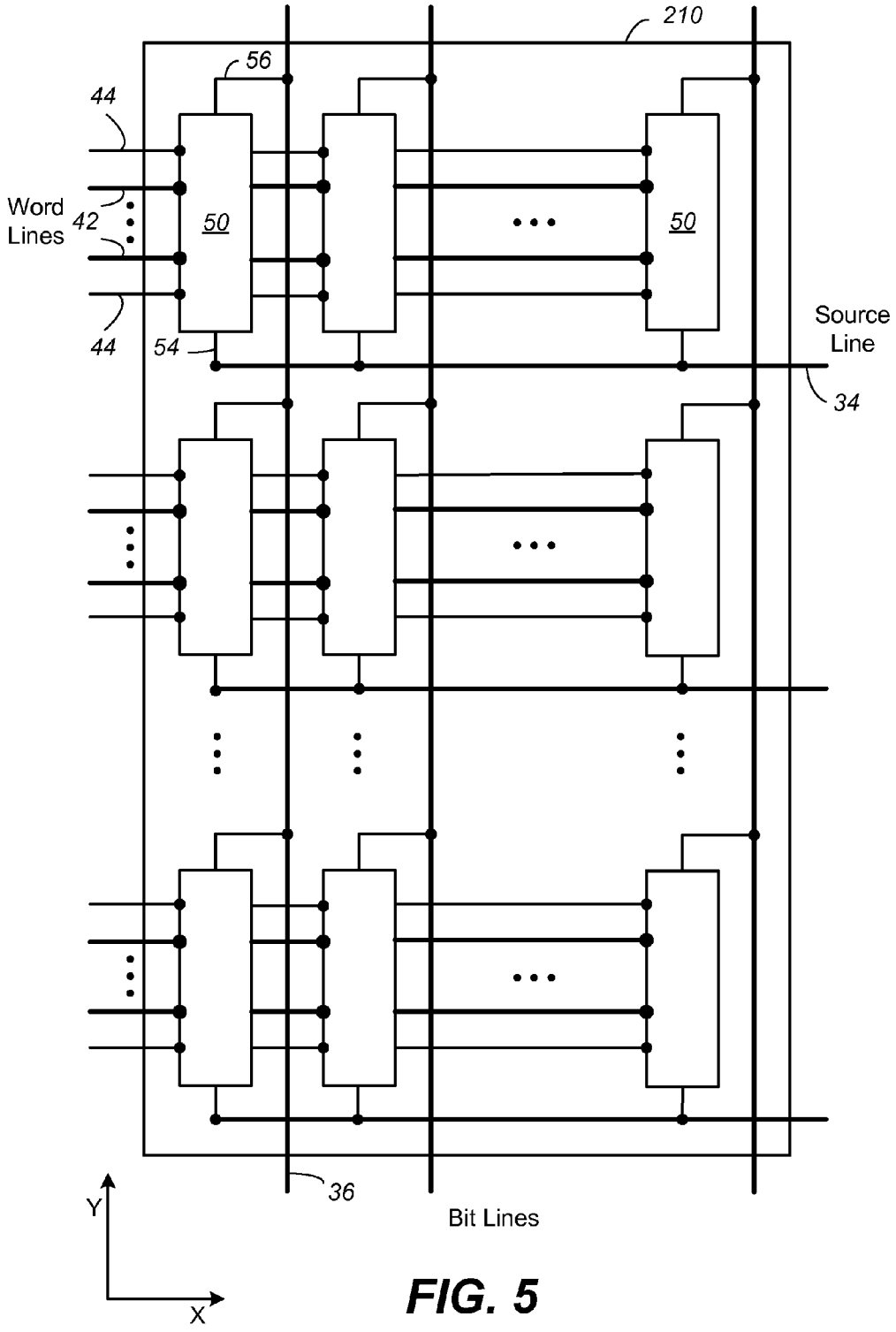

210
56
44
Word Lines
42
50
50
44
54
Source Line
34
36
Bit Lines
Y
X

FIG. 5

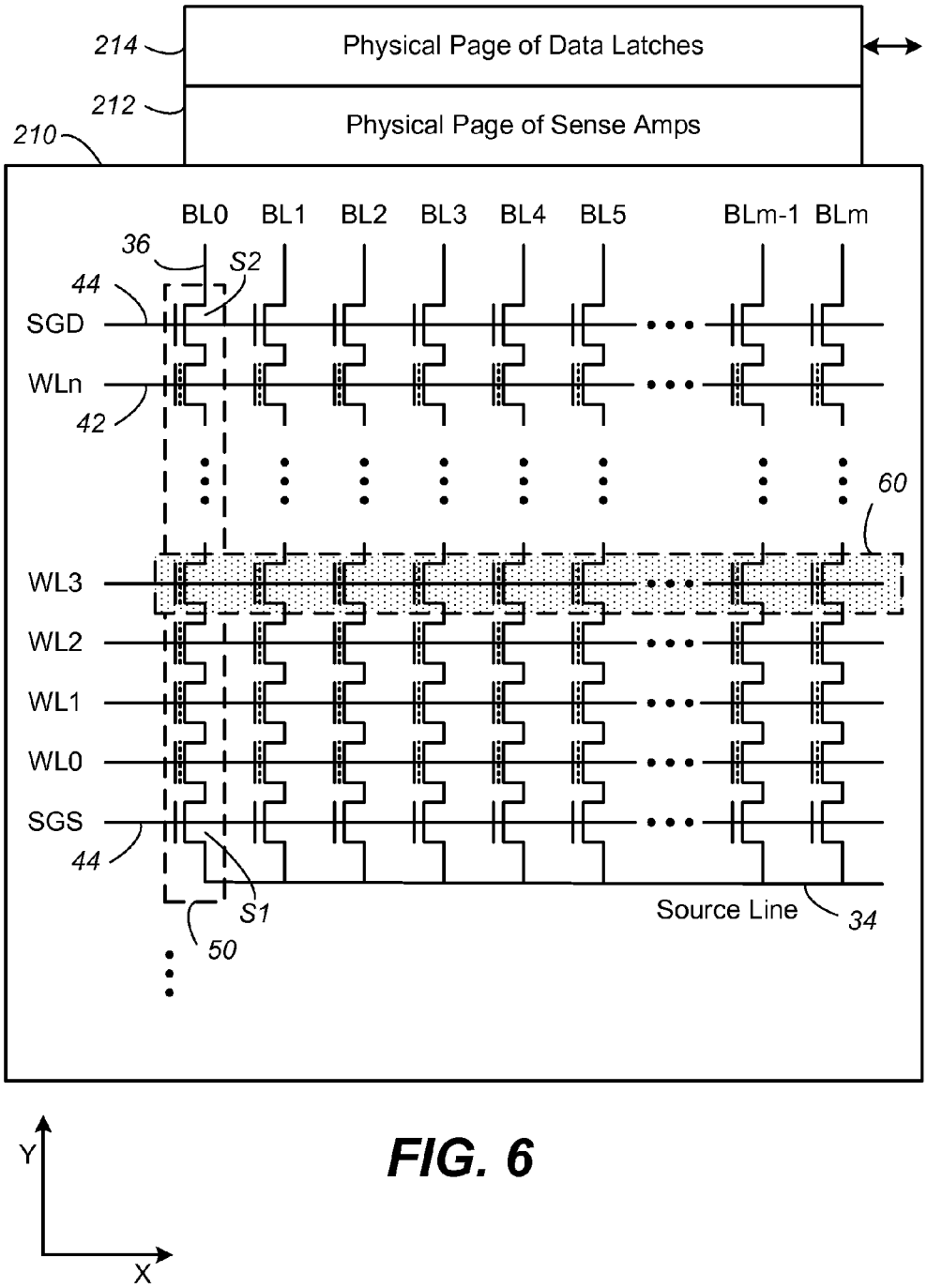
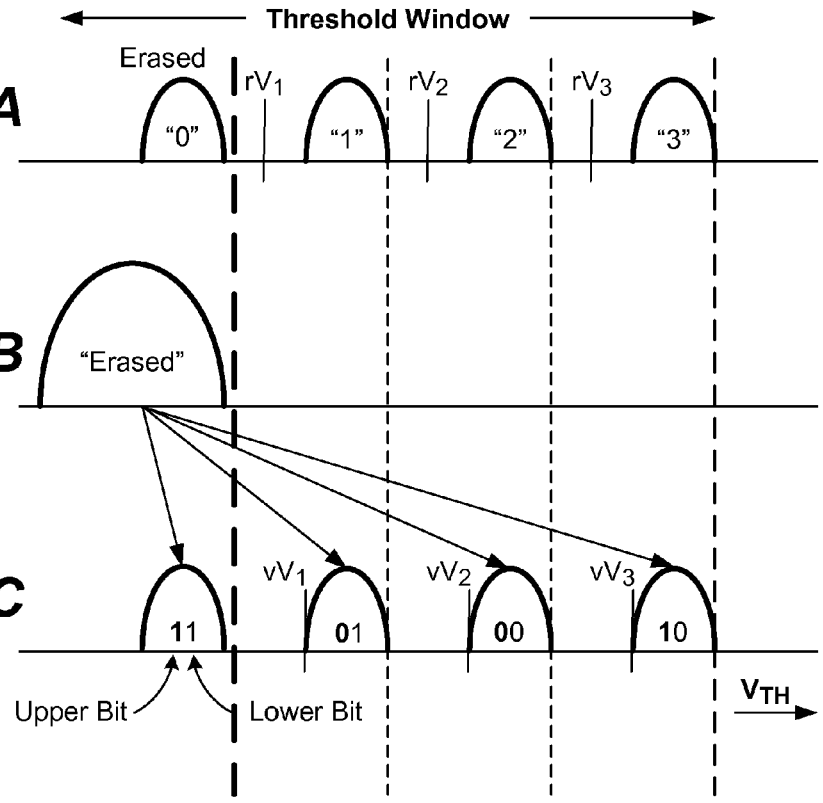

214
Physical Page of Data Latches
212
Physical Page of Sense Amps
210
BL0
BL1
BL2
BL3
BL4
BL5
BLm-1
BLm
36
S2
44
SGD
WLn
42
60
WL3
WL2
WL1
WL0
SGS
44
S1
50
Source Line
34
Y
X

FIG. 6

FIG. 7A

FIG. 7B

FIG. 7C



Programming into four states represented by a 2-bit code

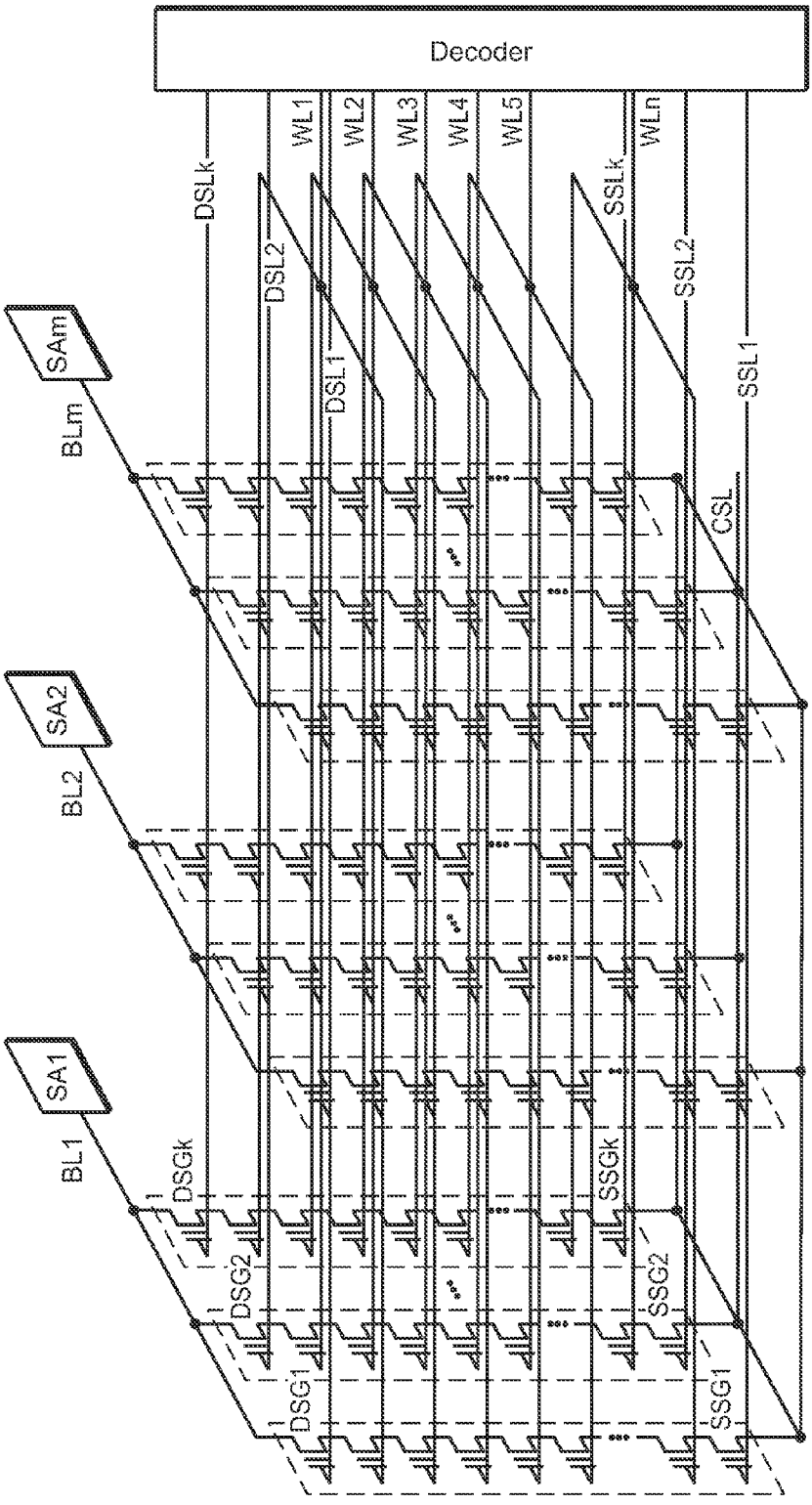

Decoder
DSLk
DSL2
DSL1
WL1
WL2
WL3
WL4
WL5
WLn
SSLk
SSL2
SSL1
CSL
SAm
BLm
SA2
BL2
SA1
BL1
DSGk
DSG2
DSG1
SSGk
SSG2
SSG1
y
x
z

FIG. 8

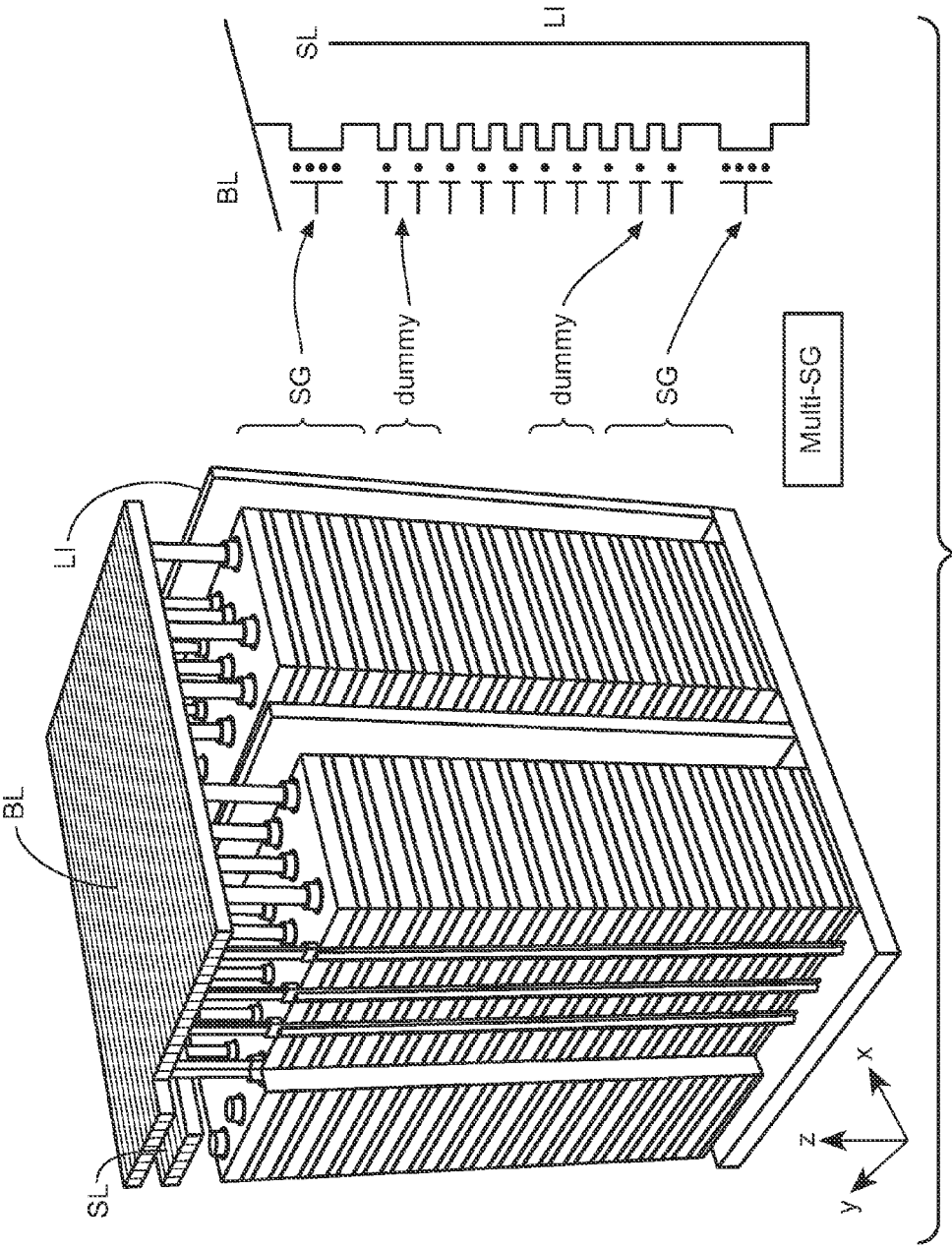


FIG. 9

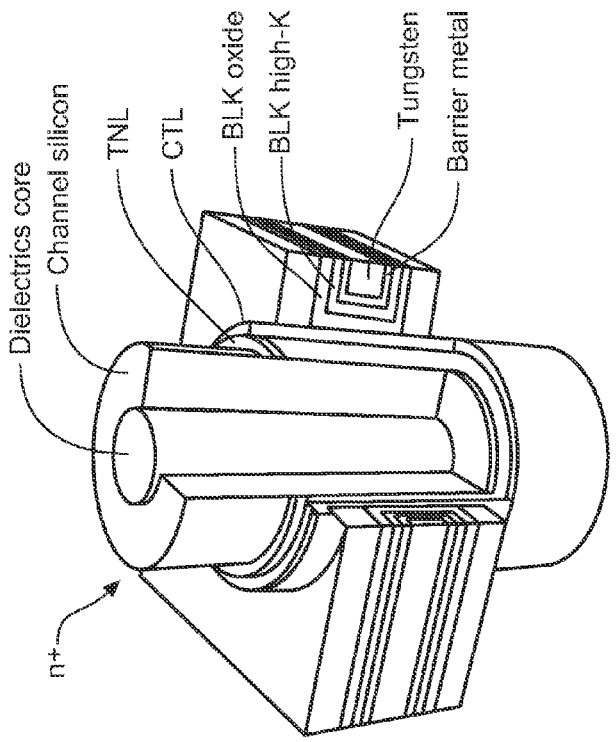

Dielectrics core
Channel silicon
TNL
CTL
BLK oxide
BLK high-K
Tungsten
Barrier metal
n+

FIG. 12

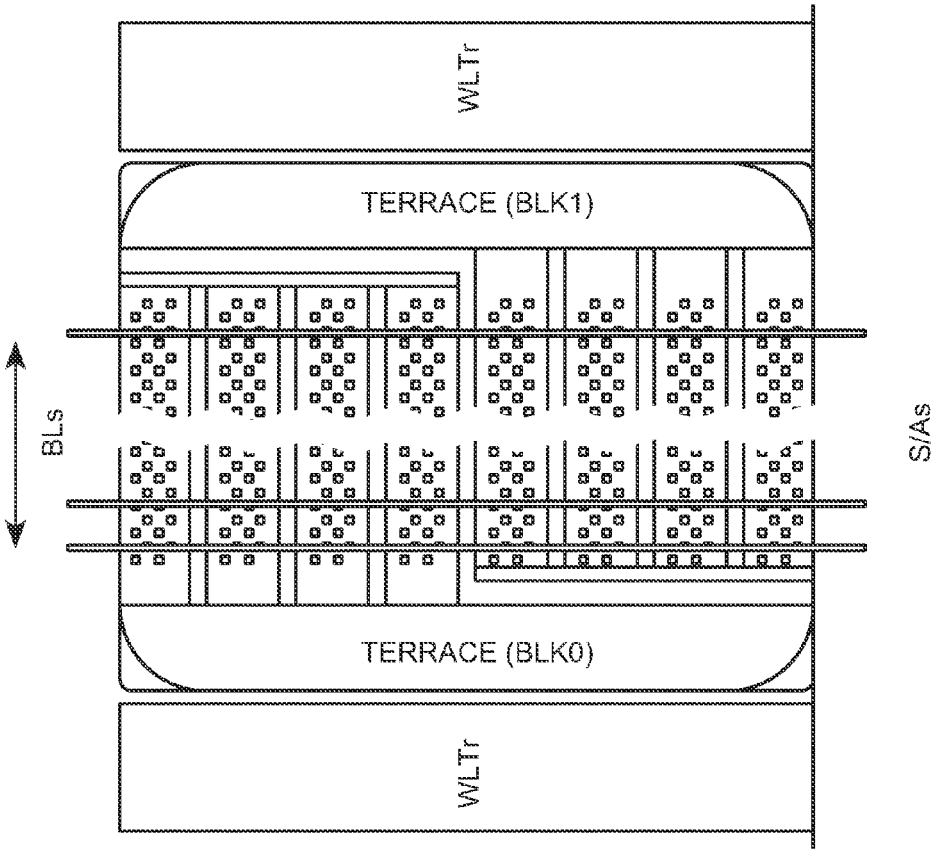

WLTr
TERRACE (BLK1)
BLs
S/As
TERRACE (BLK0)
WLTr

FIG. 10

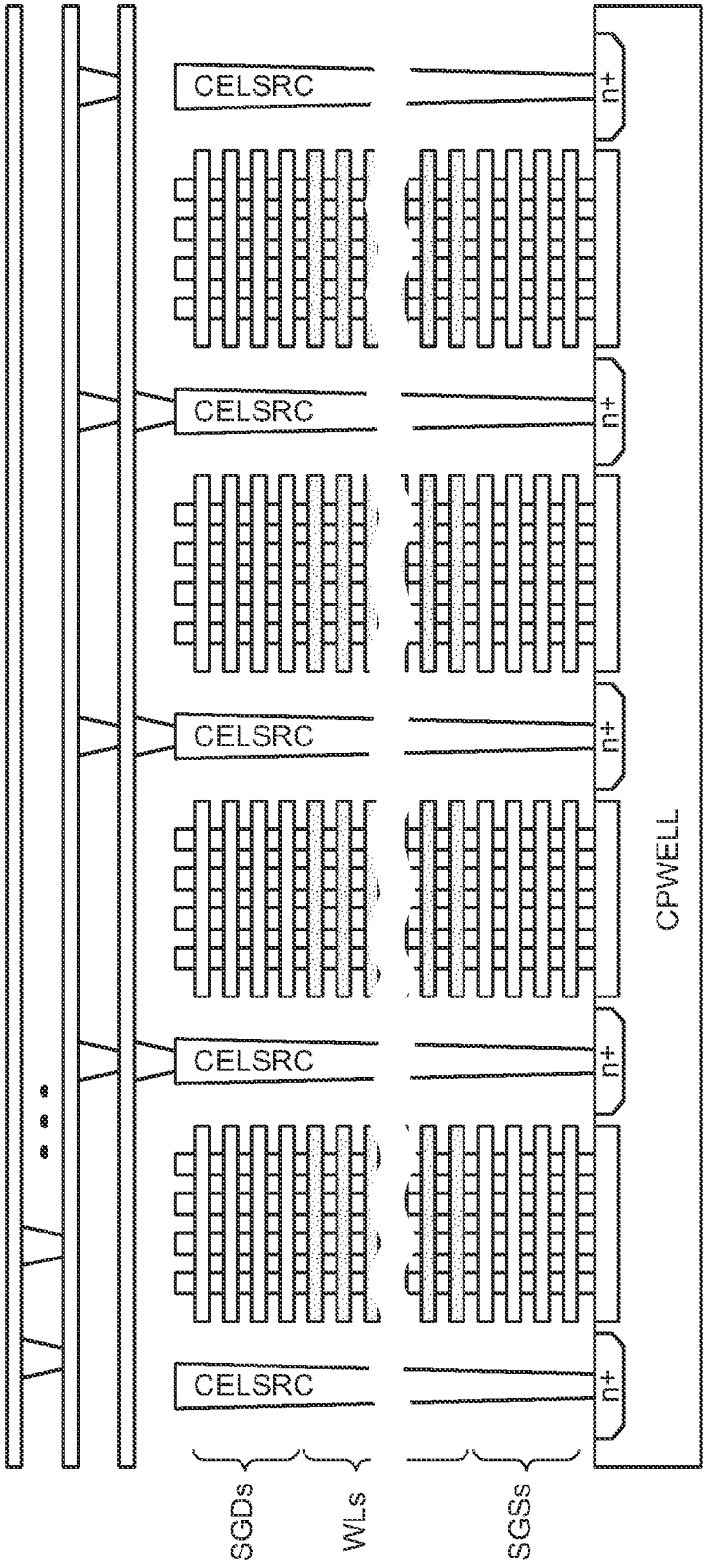

CELSRC
CELSRC
CELSRC
CELSRC
CELSRC
n+
n+
n+
n+
n+
CPWELL
SGDs
WLs
SGSs

FIG. 11

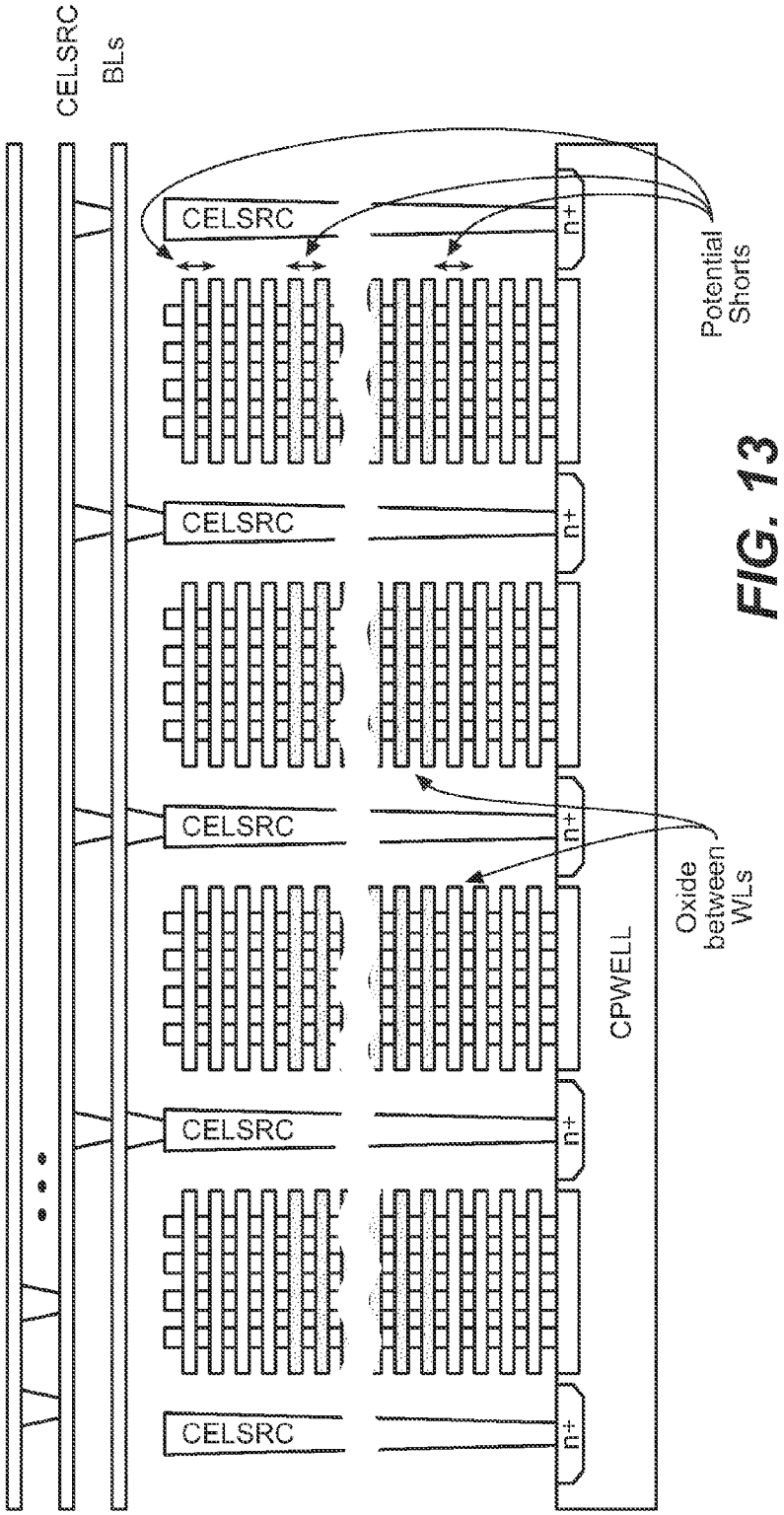

CELSRC
BLs
CELSRC
CELSRC
CELSRC
CELSRC
CELSRC
n+
n+
n+
n+
n+
Potential Shorts
Oxide between WLs
CPWELL

FIG. 13

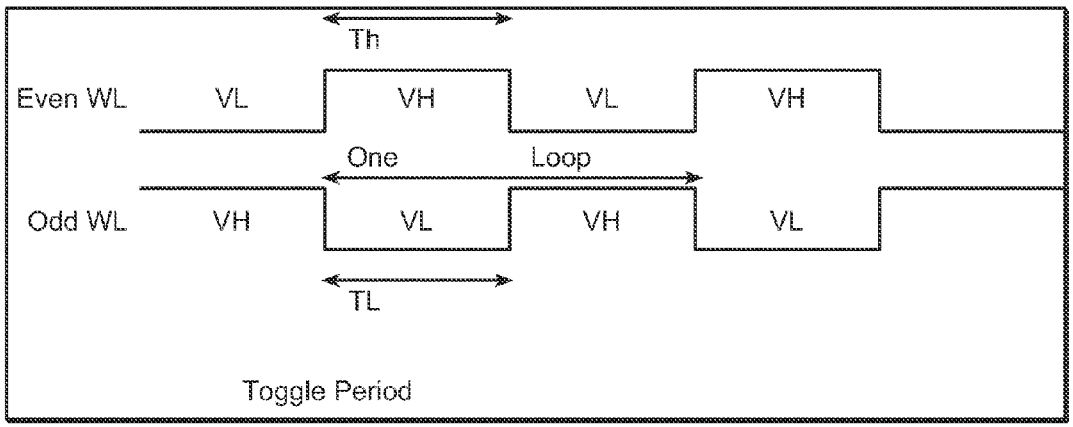

Th
Even WL
VL
VH
VL
VH
One
Loop
Odd WL
VH
VL
VH
VL
TL
Toggle Period

FIG. 14

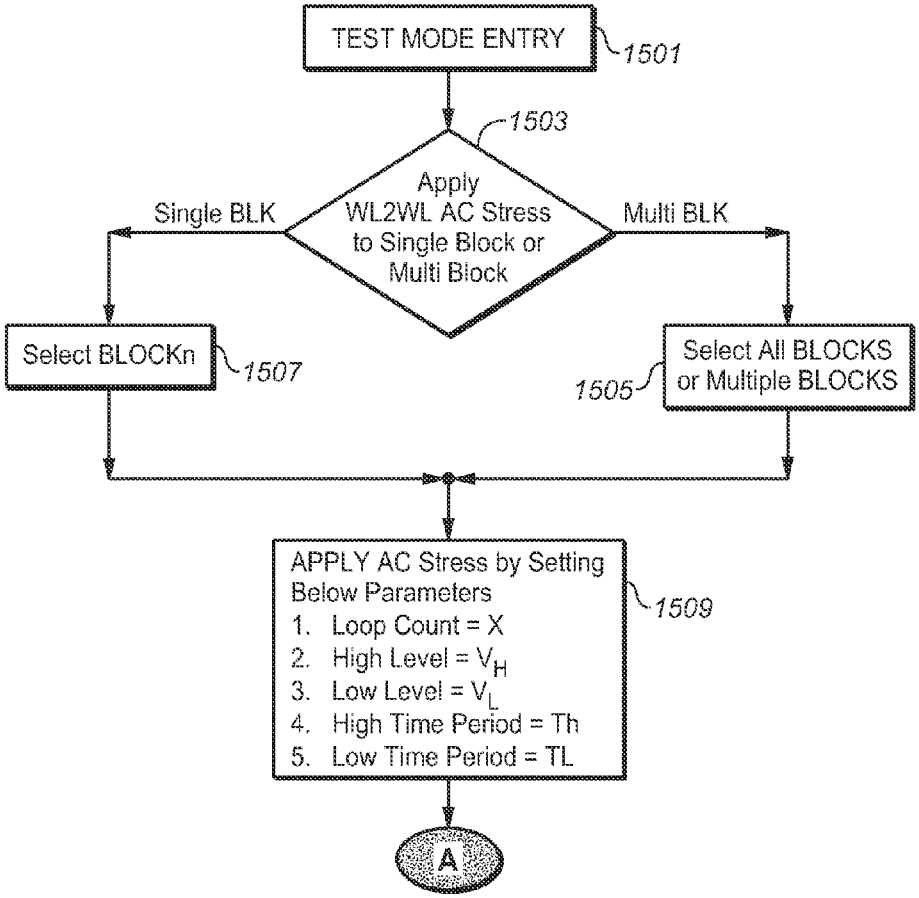

TEST MODE ENTRY
1501
1503
Apply
WL2WL AC Stress
to Single Block or
Multi Block
Single BLK
Multi BLK
Select BLOCKn
1507
1505
Select All BLOCKS
or Multiple BLOCKS
APPLY AC Stress by Setting
Below Parameters
1. Loop Count = X
2. High Level = V_H
3. Low Level = V_L
4. High Time Period = Th
5. Low Time Period = TL
1509
A

FIG. 15

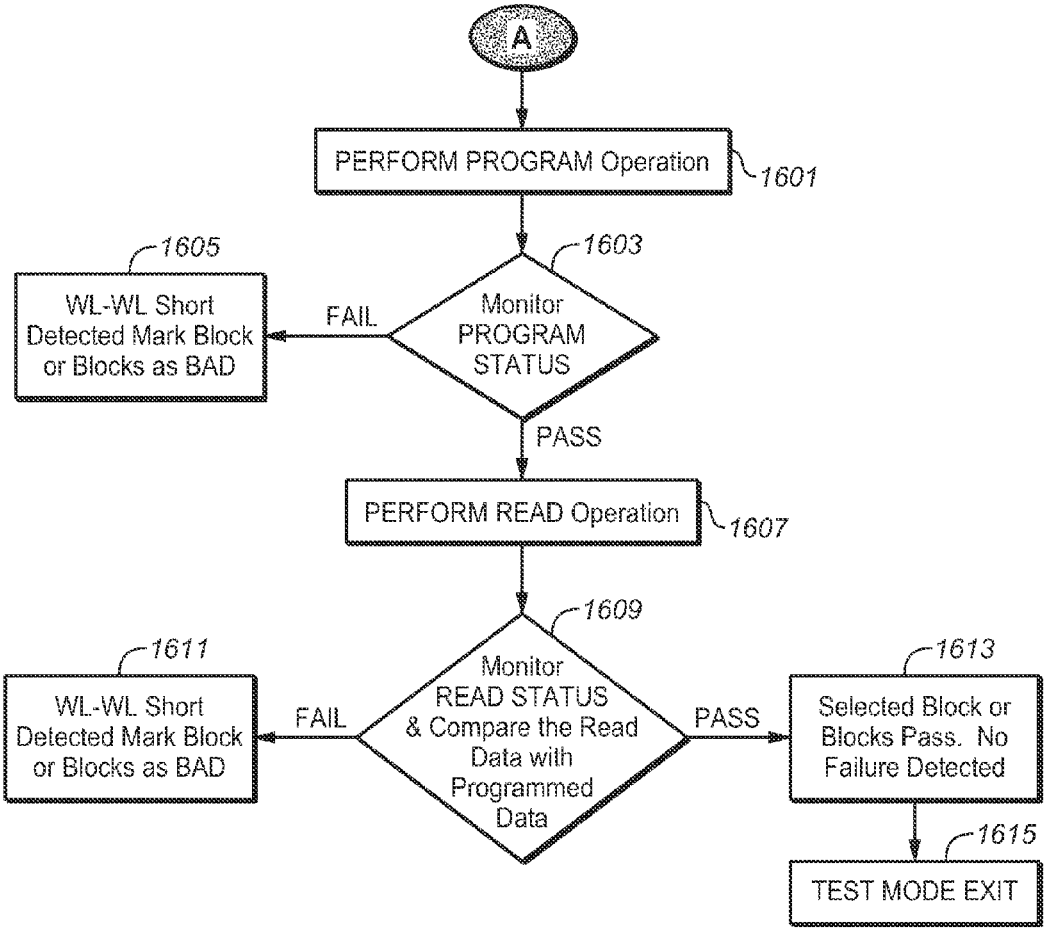

A
PERFORM PROGRAM Operation
1601
1603
Monitor PROGRAM STATUS
FAIL
1605
WL-WL Short Detected Mark Block or Blocks as BAD
PASS
PERFORM READ Operation
1607
1609
Monitor READ STATUS & Compare the Read Data with Programmed Data
FAIL
1611
WL-WL Short Detected Mark Block or Blocks as BAD
PASS
1613
Selected Block or Blocks Pass. No Failure Detected
1615
TEST MODE EXIT

FIG. 16

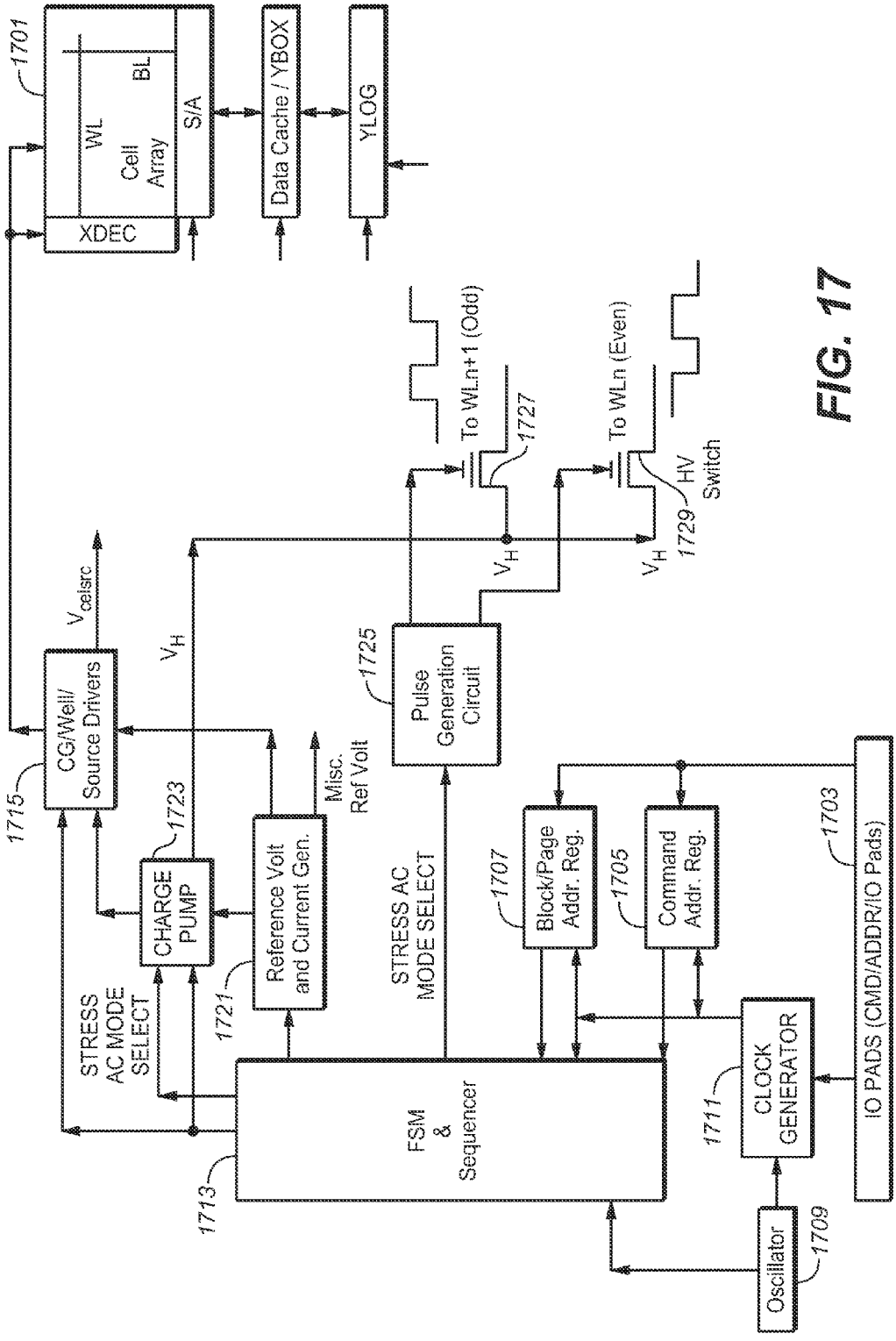

1701
XDEC
Cell Array
WL
BL
S/A
Data Cache / YBOX
YLOG
To WLn+1 (Odd)
To WLn (Even)
1727
HV Switch
1729
V_H
V_{celsrc}
CG/Well/ Source Drivers
1715
CHARGE PUMP
1723
Reference Volt and Current Gen.
1721
Misc. Ref Volt
Pulse Generation Circuit
1725
STRESS AC MODE SELECT
STRESS AC MODE SELECT
1707
Block/Page Addr. Reg.
1705
Command Addr. Reg.
1703
IO PADS (CMD/ADDR/IO Pads)
FSM & Sequencer
1713
CLOCK GENERATOR
1711
Oscillatr
1709

FIG. 17

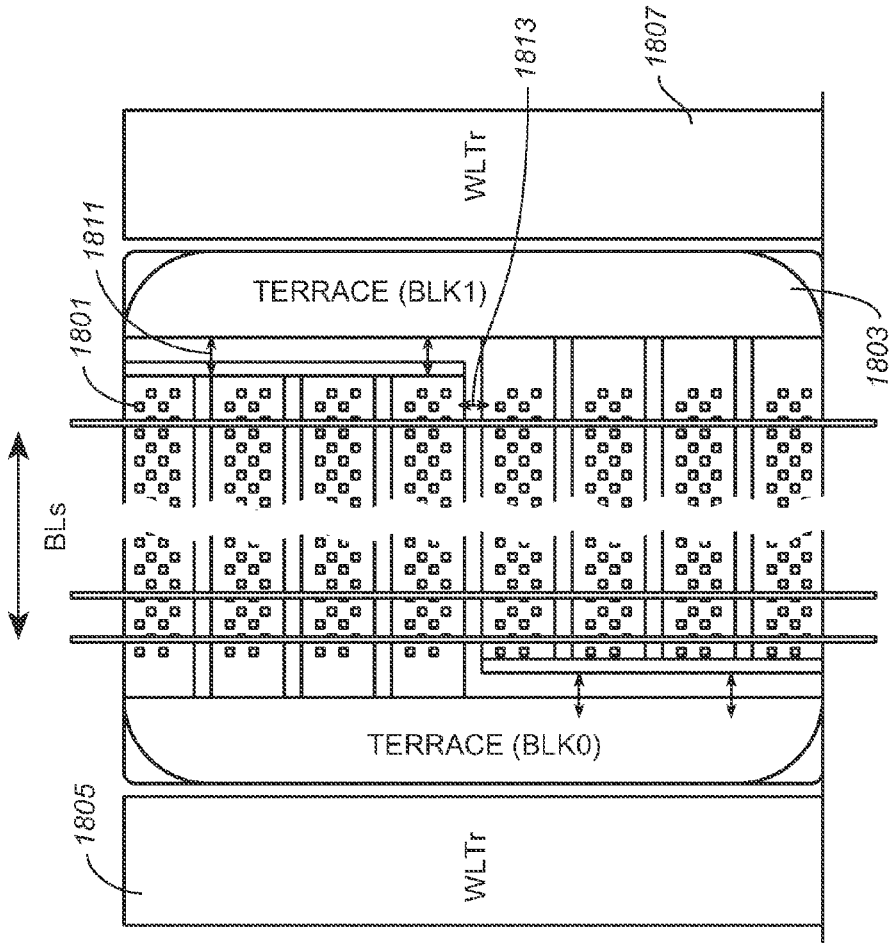

1813
1807
WLTr
1811
TERRACE (BLK1)
1801
1803
BLs
S/A (8kB)
TERRACE (BLK0)
1805
WLTr

FIG. 18

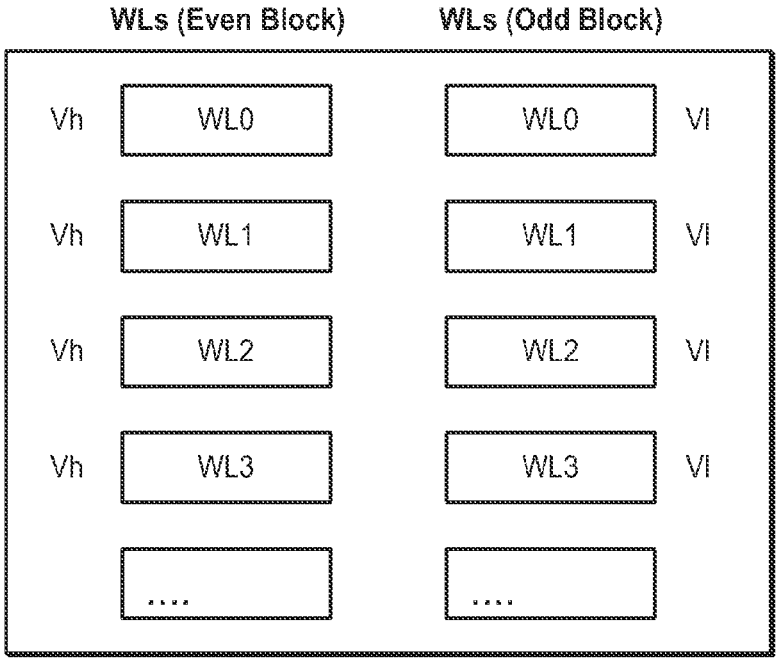

WLs (Even Block)
WLs (Odd Block)
Vh
WL0
WL0
Vl
Vh
WL1
WL1
Vl
Vh
WL2
WL2
Vl
Vh
WL3
WL3
Vl
....
....

FIG. 19

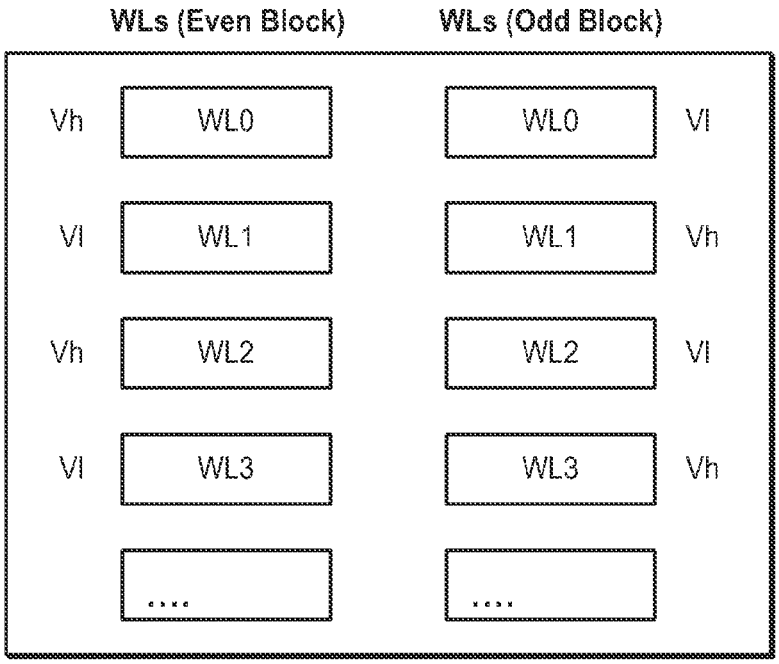

WLs (Even Block)
WLs (Odd Block)
Vh
WL0
WL0
Vl
Vl
WL1
WL1
Vh
Vh
WL2
WL2
Vl
Vl
WL3
WL3
Vh
....
....

FIG. 20

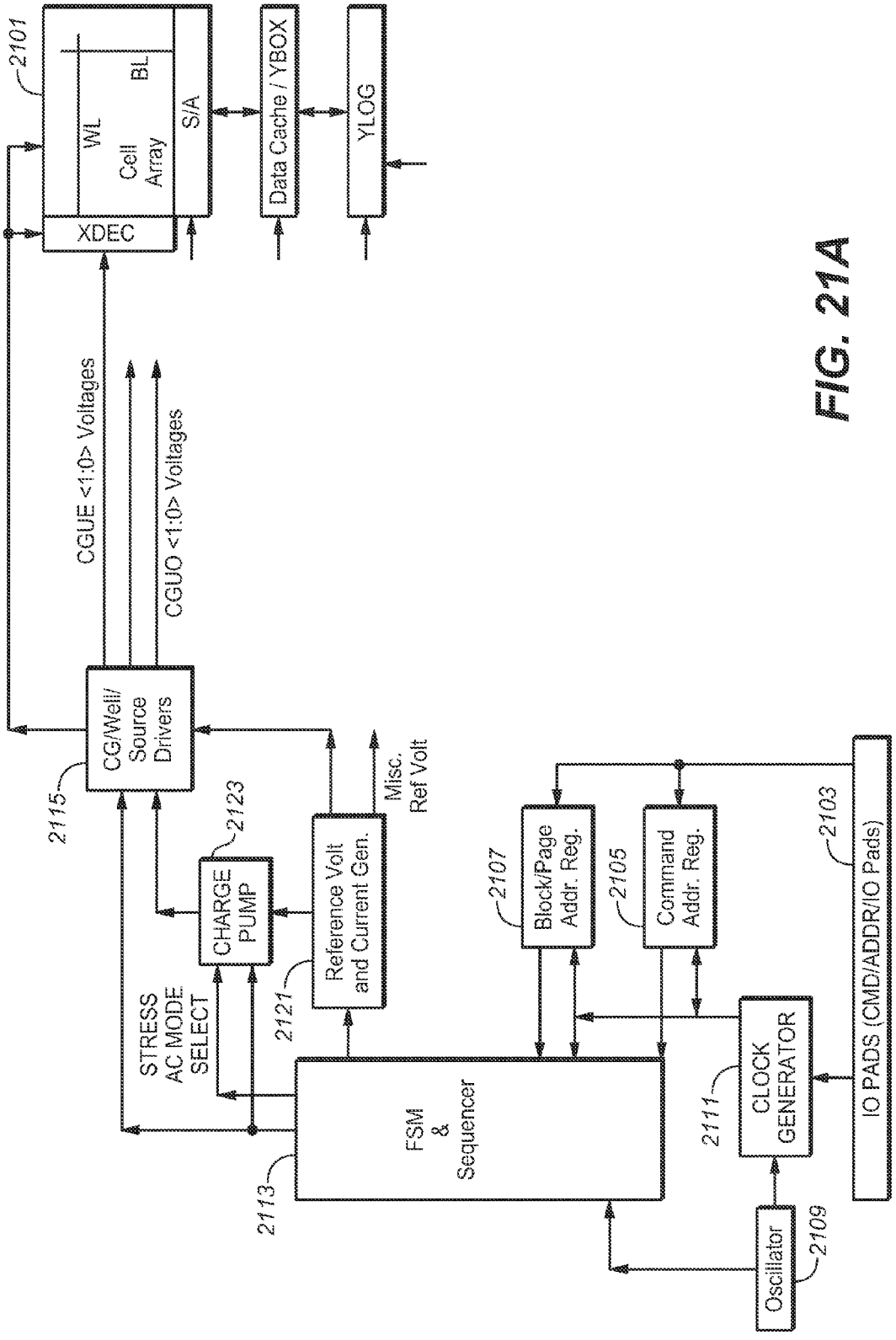

2101
WL
BL
Cell Array
S/A
XDEC
Data Cache / YBOX
YLOG
CGUE <1:0> Voltages
CGUO <1:0> Voltages
CG/Well/ Source Drivers
2115
CHARGE PUMP
2123
Reference Volt and Current Gen.
2121
Misc. Ref Volt
STRESS AC MODE SELECT
FSM & Sequencer
2113
Block/Page Addr. Reg.
2107
Command Addr. Reg.
2105
CLOCK GENERATOR
2111
Oscillator
2109
IO PADS (CMD/ADDR/IO Pads)
2103

FIG. 21A

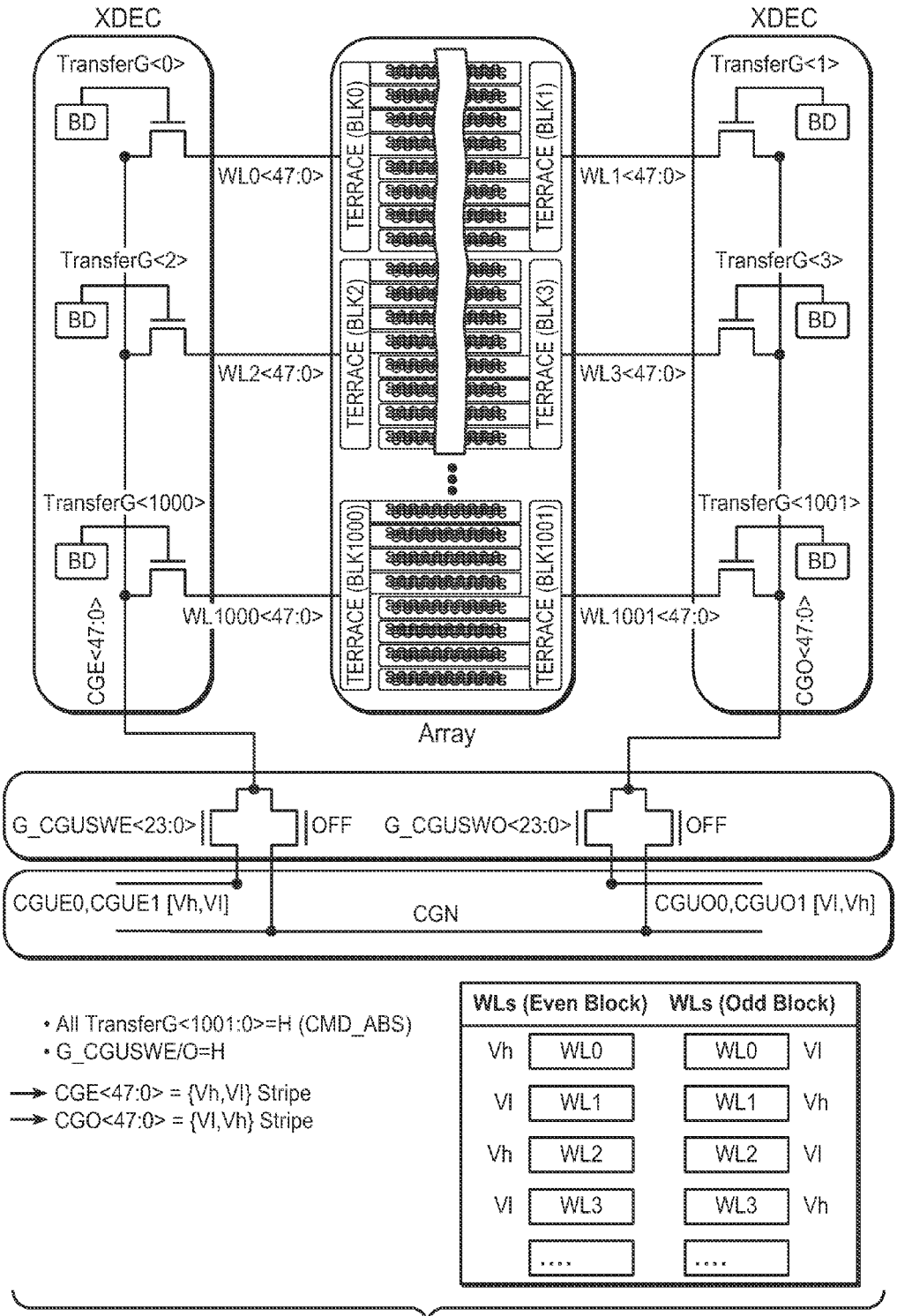

XDEC
XDEC
TransferG<0>
TransferG<1>
BD
BD
WL0<47:0>
WL1<47:0>
TERRACE (BLK0)
TERRACE (BLK1)
TransferG<2>
TransferG<3>
BD
BD
WL2<47:0>
WL3<47:0>
TERRACE (BLK2)
TERRACE (BLK3)
TransferG<1000>
TransferG<1001>
BD
BD
WL1000<47:0>
WL1001<47:0>
TERRACE (BLK1000)
TERRACE (BLK1001)
CGE<47:0>
CGO<47:0>
Array
G_CGUSWE<23:0>
OFF
G_CGUSWO<23:0>
OFF
CGUE0,CGUE1 [Vh,Vl]
CGN
CGUO0,CGUO1 [Vl,Vh]
• All TransferG<1001:0>=H (CMD_ABS)
• G_CGUSWE/O=H
CGE<47:0> = {Vh,Vl} Stripe
CGO<47:0> = {Vl,Vh} Stripe
WLs (Even Block)
WLs (Odd Block)
Vh
WL0
WL0
Vl
Vl
WL1
WL1
Vh
Vh
WL2
WL2
Vl
Vl
WL3
WL3
Vh
....
....

FIG. 21B

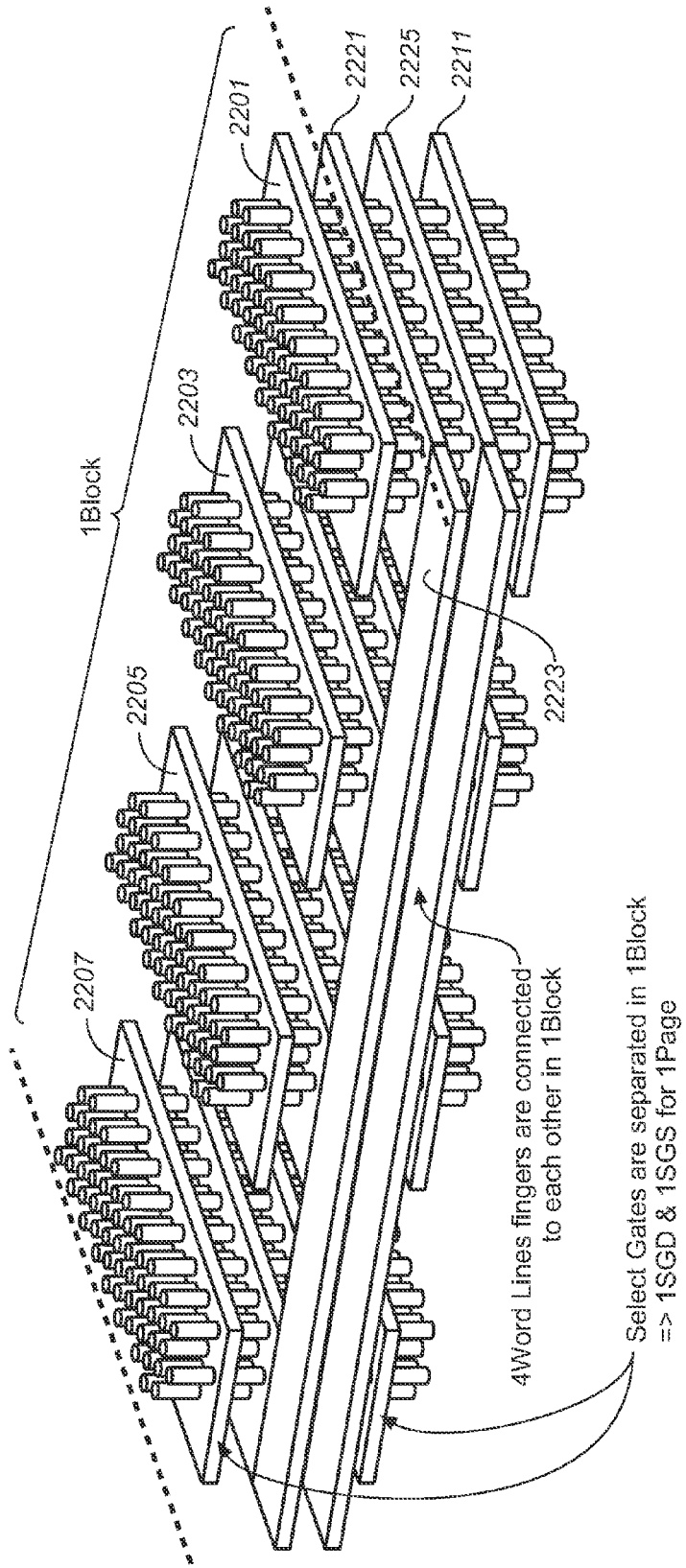

1Block
2201
2221
2225
2211
2203
2205
2223
2207
4Word Lines fingers are connected to each other in 1Block
Select Gates are separated in 1Block
=> 1SGD & 1SGS for 1Page

FIG. 22

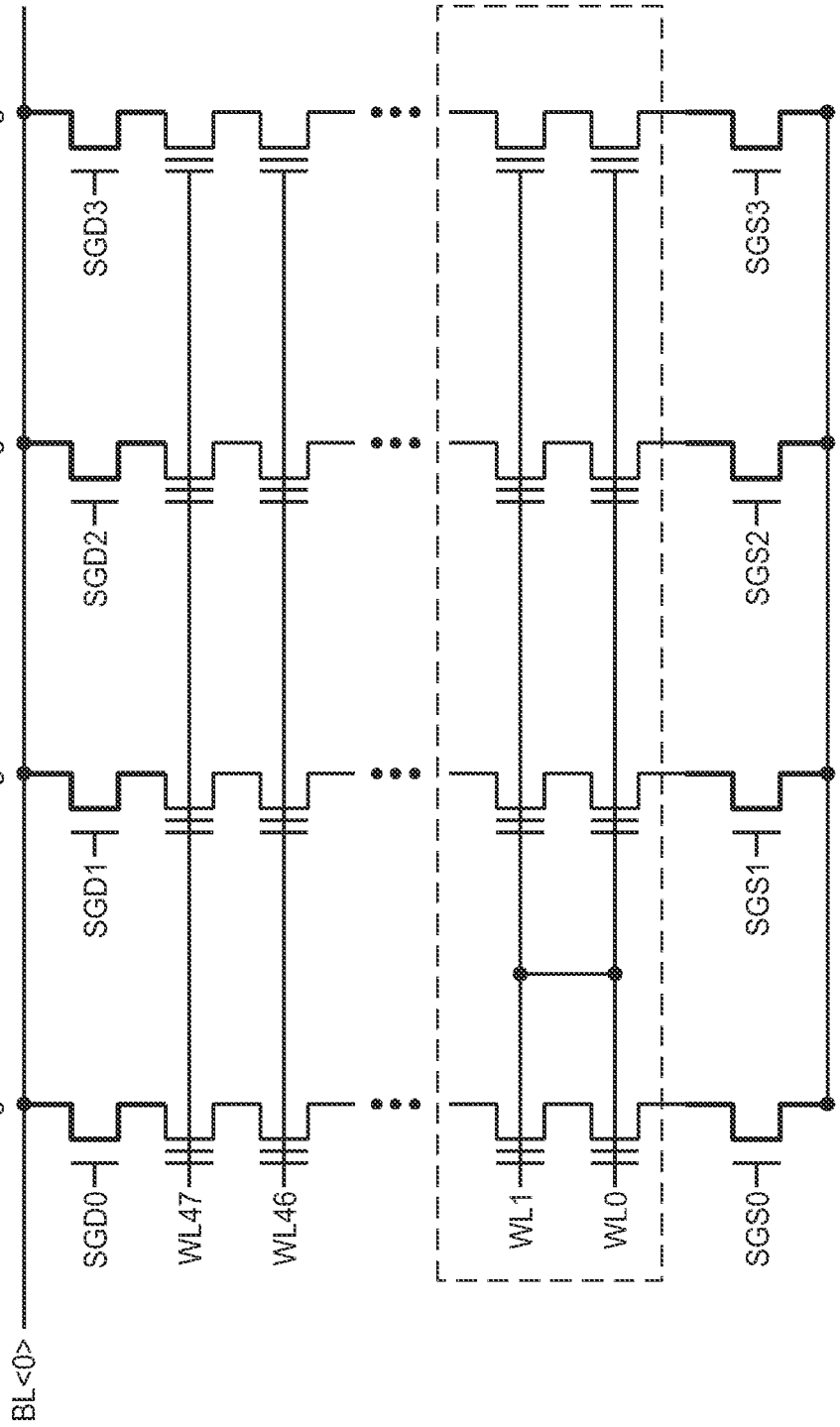

Finger 3
Finger 2
Finger 1
Finger 0
SGD3
SGD2
SGD1
SGD0
WL47
WL46
WL1
WL0
SGS3
SGS2
SGS1
SGS0
BL<0>

FIG. 23

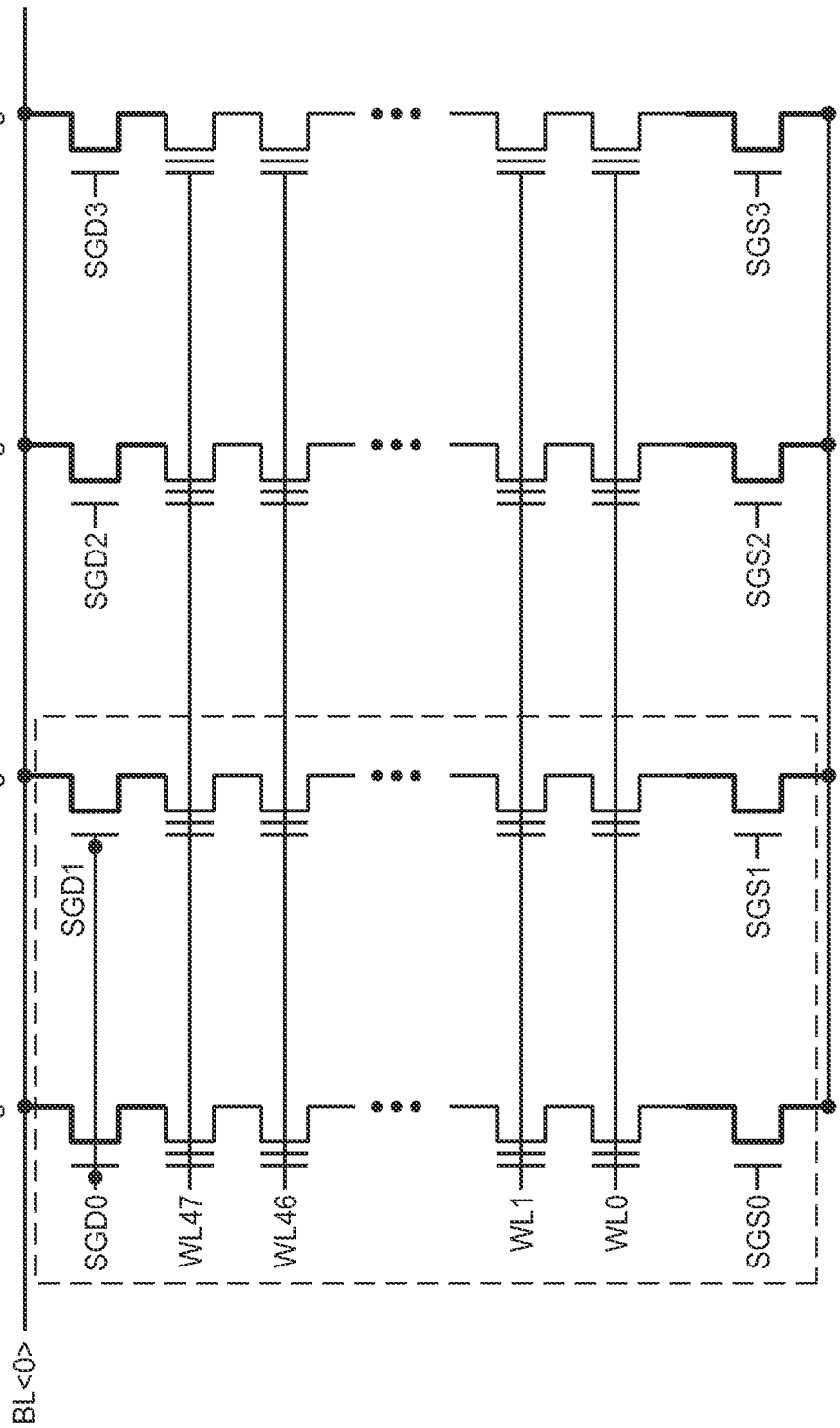

Finger 0
Finger 1
Finger 2
Finger 3
BL<0>
SGD0
SGD1
SGD2
SGD3
WL47
WL46
WL1
WL0
SGS0
SGS1
SGS2
SGS3

FIG. 24

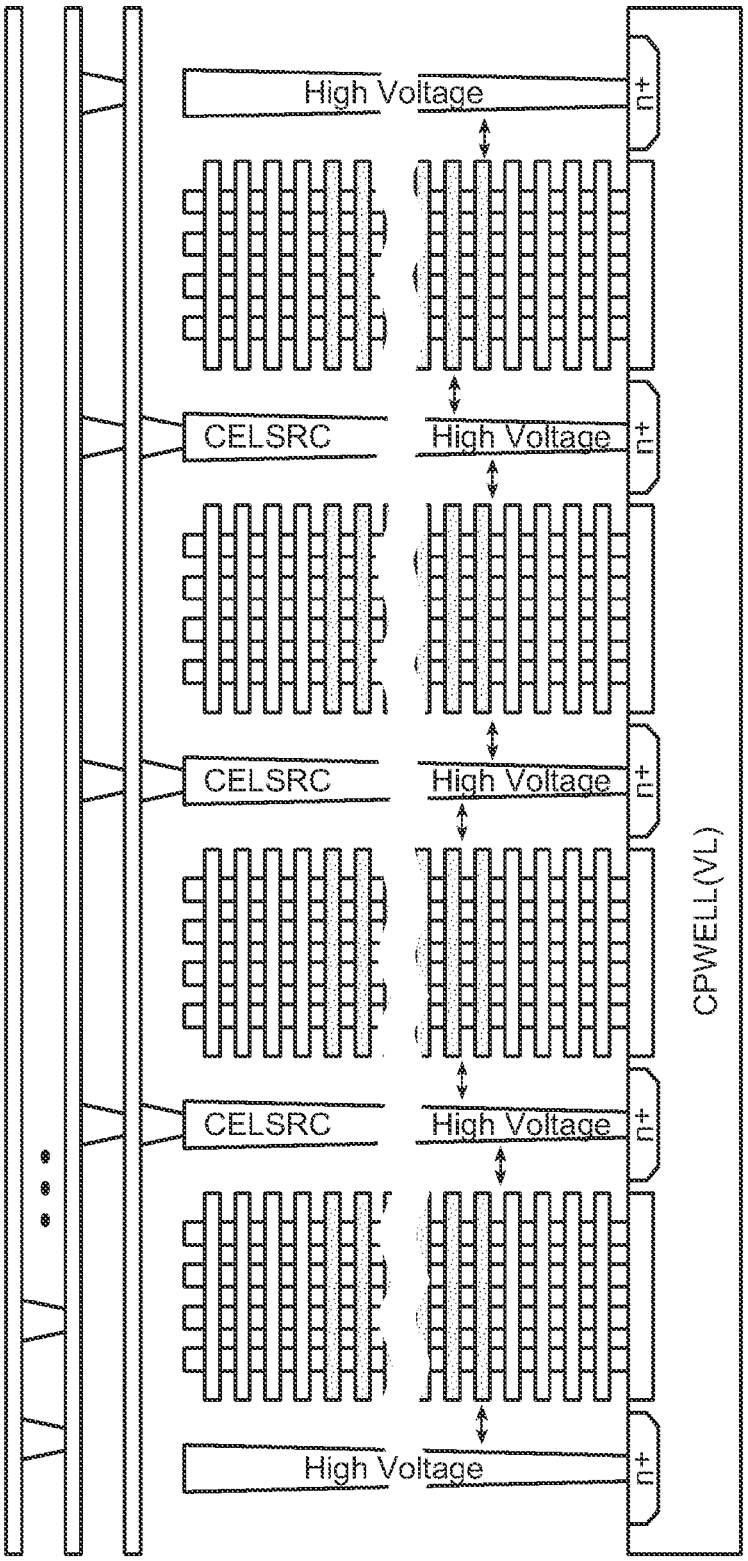

High Voltage
n+
CELSRC
High Voltage
n+
CELSRC
High Voltage
n+
CELSRC
High Voltage
n+
High Voltage
n+
CPWELL(VL)

FIG. 25

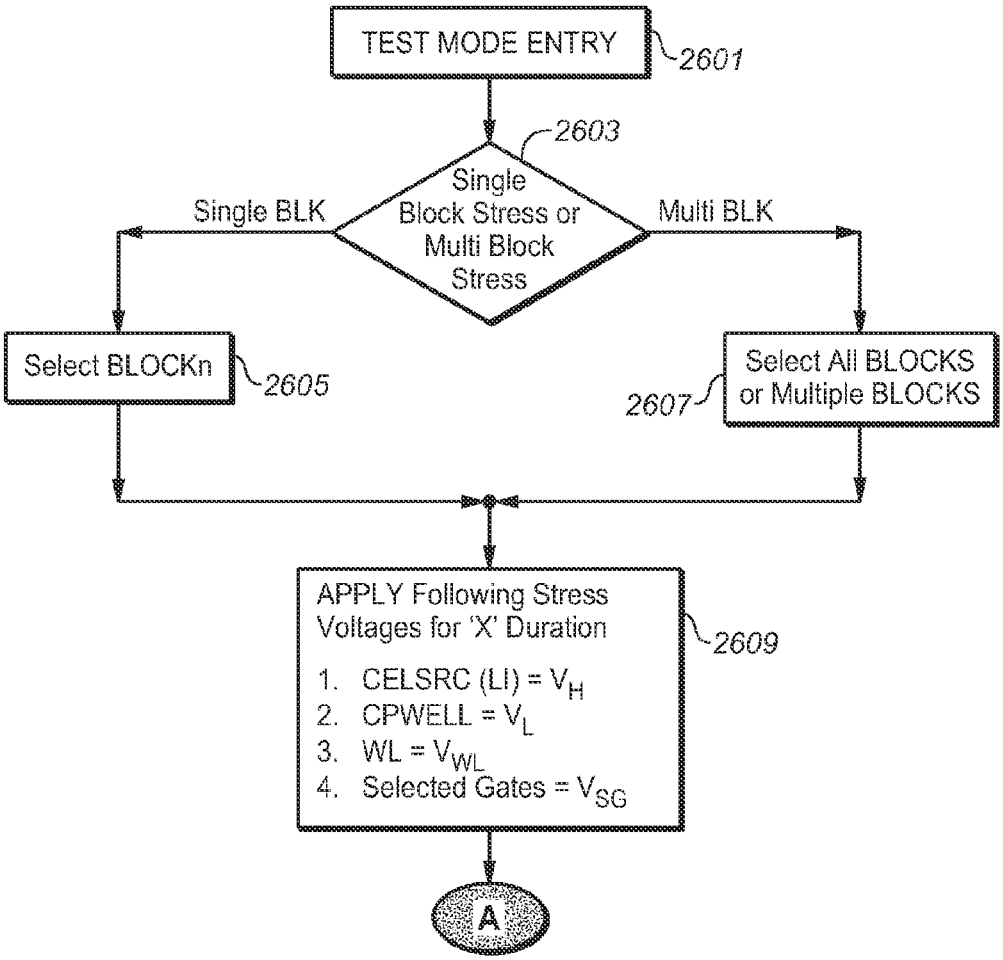

TEST MODE ENTRY
2601
2603
Single Block Stress or Multi Block Stress
Single BLK
Multi BLK
Select BLOCKn
2605
2607
Select All BLOCKS or Multiple BLOCKS
APPLY Following Stress Voltages for 'X' Duration
2609
1. CELSRC (LI) = V_H
2. CPWELL = V_L
3. WL = V_{WL}
4. Selected Gates = V_{SG}
A

FIG. 26

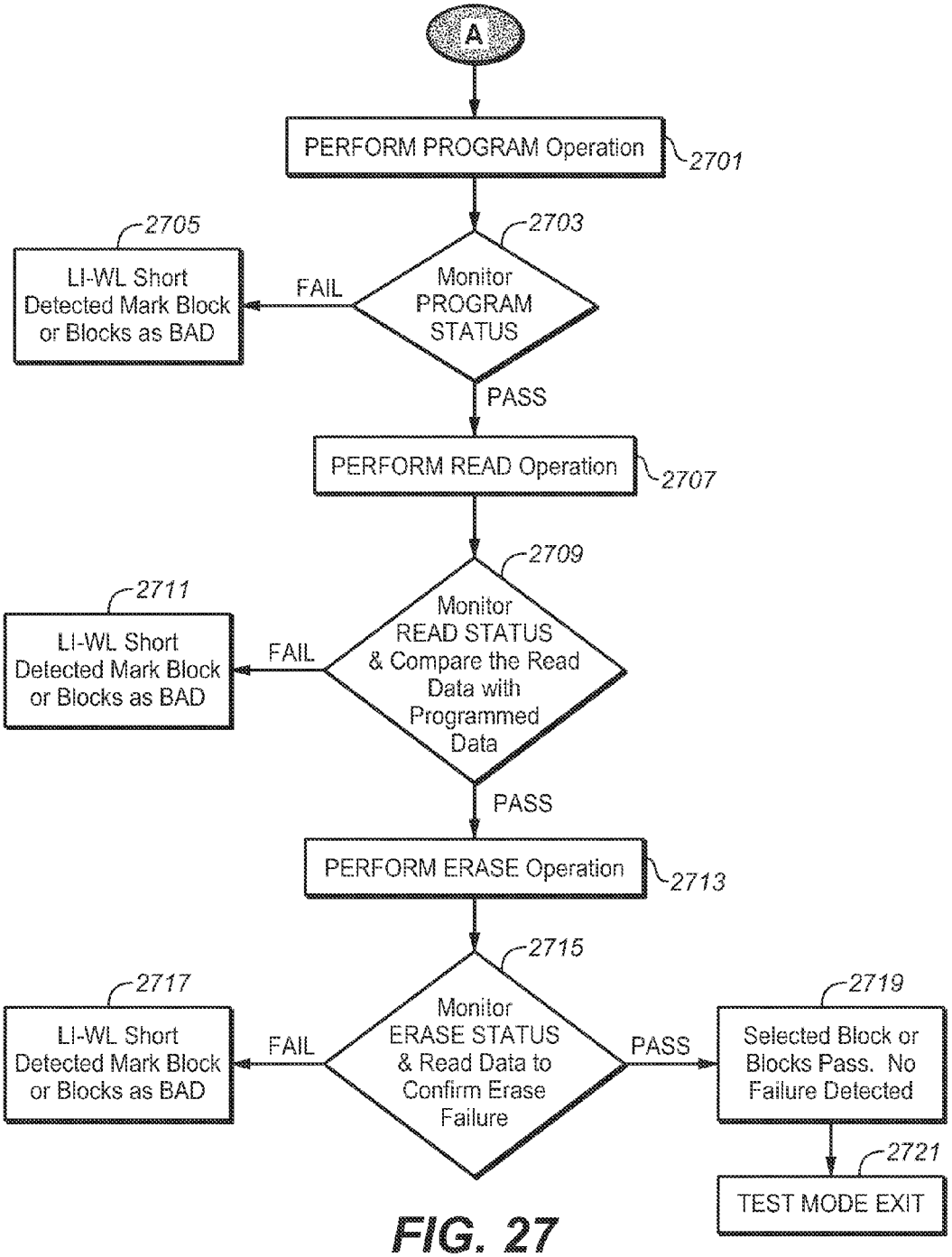

A
PERFORM PROGRAM Operation
2701
2703
Monitor PROGRAM STATUS
FAIL
2705
LI-WL Short Detected Mark Block or Blocks as BAD
PASS
PERFORM READ Operation
2707
2709
Monitor READ STATUS & Compare the Read Data with Programmed Data
FAIL
2711
LI-WL Short Detected Mark Block or Blocks as BAD
PASS
PERFORM ERASE Operation
2713
2715
Monitor ERASE STATUS & Read Data to Confirm Erase Failure
FAIL
2717
LI-WL Short Detected Mark Block or Blocks as BAD
PASS
2719
Selected Block or Blocks Pass. No Failure Detected
2721
TEST MODE EXIT

FIG. 27

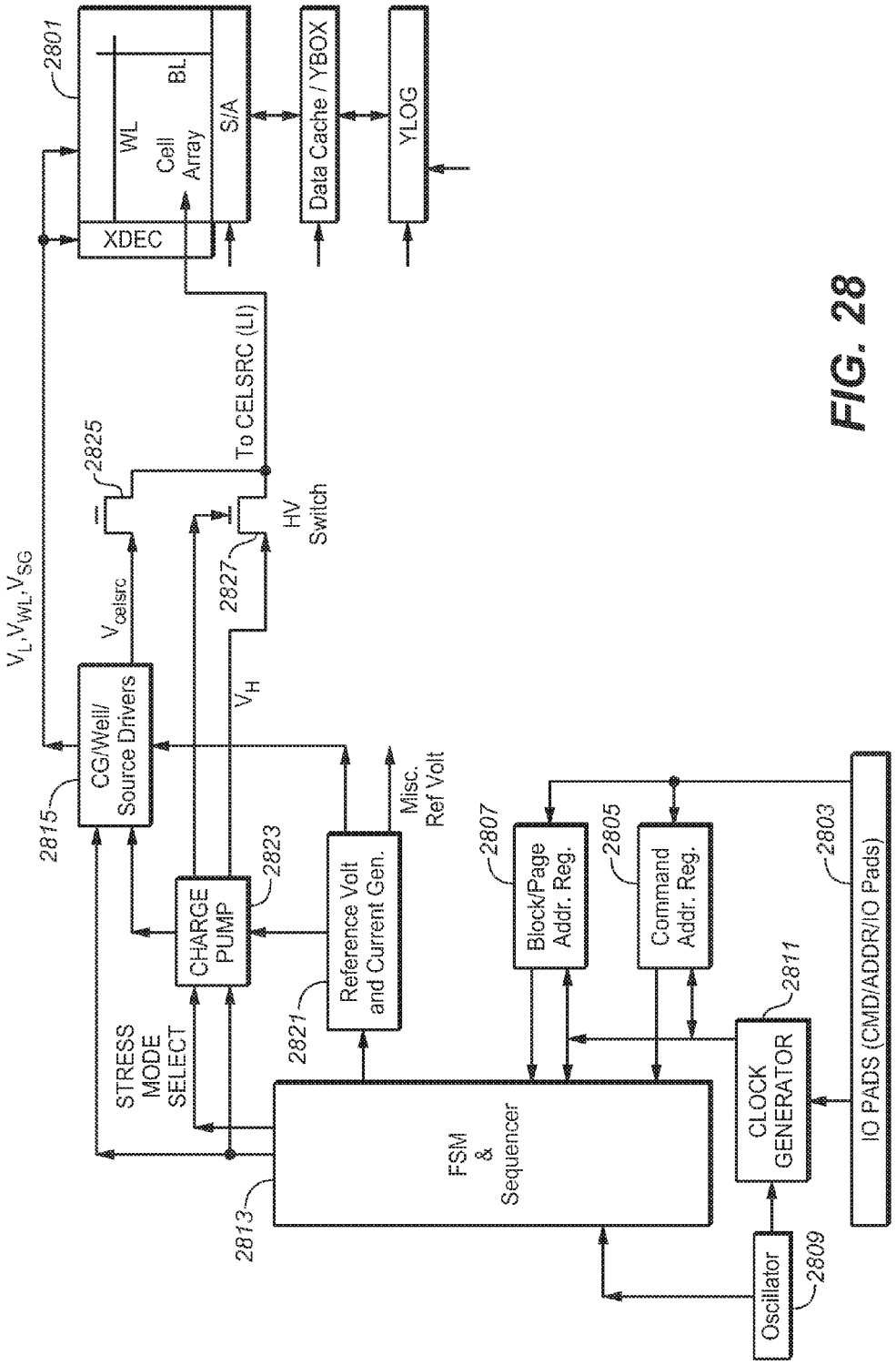

2801
WL
Cell Array
BL
S/A
XDEC
Data Cache / YBOX
YLOG
To CELSRC (LI)
2825
HV Switch
2827
V_{celsrc}
V_H
V_L, V_{WL}, V_{SG}
2815
CG/Well/ Source Drivers
CHARGE PUMP
2823
Reference Volt and Current Gen.
Misc. Ref Volt
2821
STRESS MODE SELECT
2807
Block/Page Addr. Reg.
2805
Command Addr. Reg.
2803
IO PADS (CMD/ADDR/IO Pads)
2811
CLOCK GENERATOR
FSM & Sequencer
2813
Oscillator
2809

FIG. 28

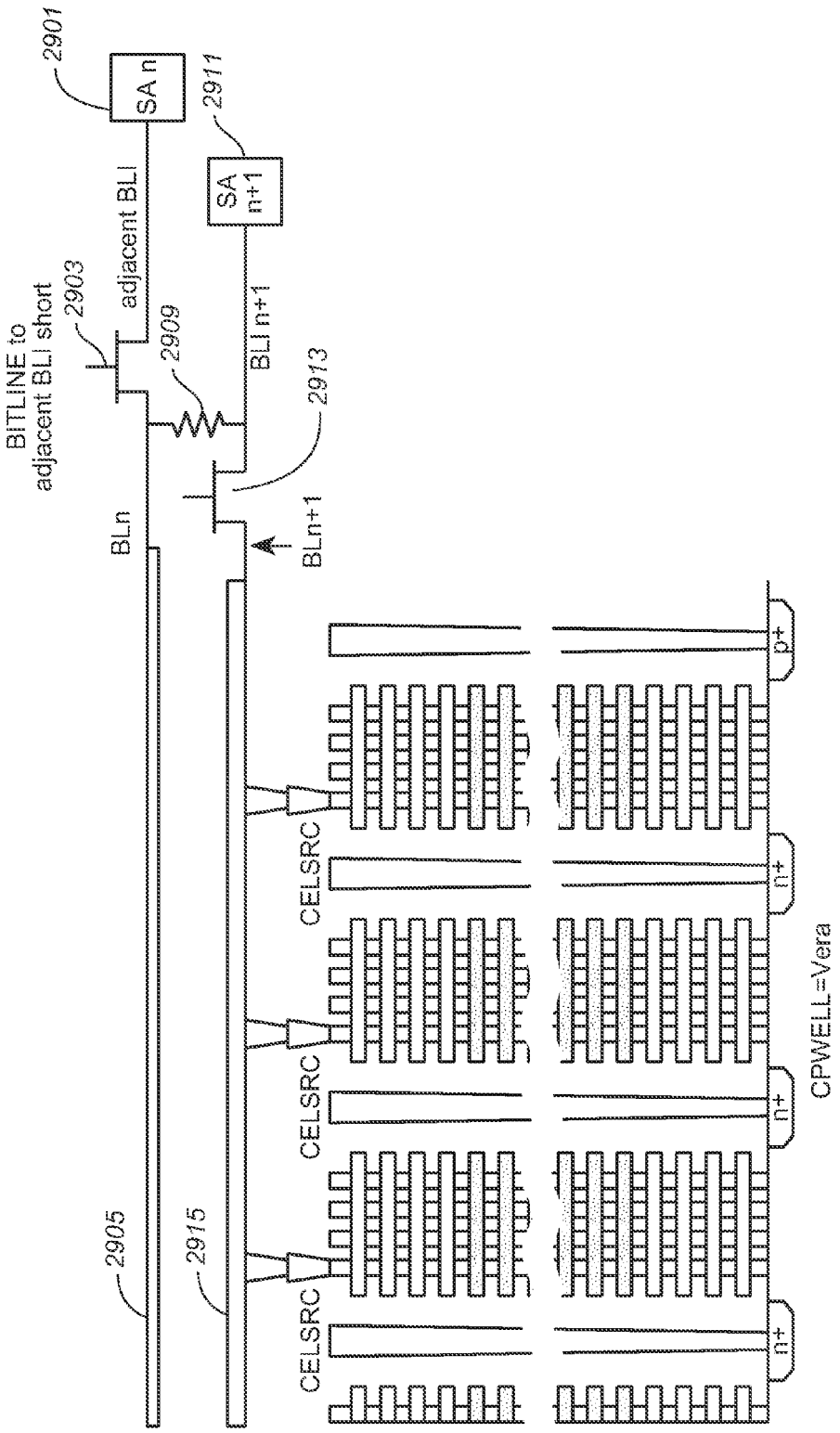

2901
SA n
2911
SA n+1
adjacent BLI
BLI n+1
BITLINE to adjacent BLI short
2903
2909
2913
BLn
BLn+1
2905
2915
CELSRC
CELSRC
CELSRC
n+
n+
n+
p+
CPWELL=Vera

FIG. 29

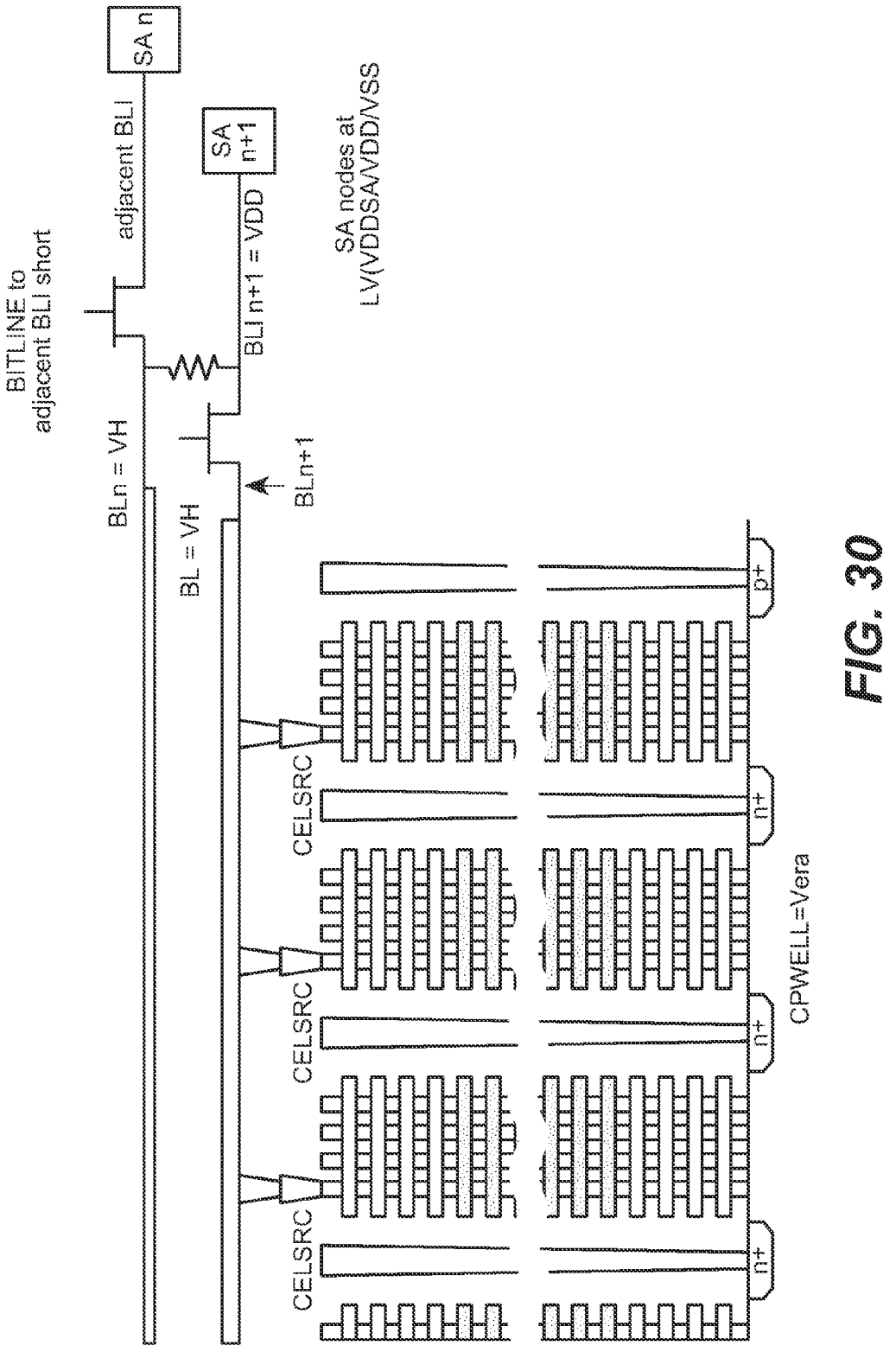

BITLINE to
adjacent BLI short
adjacent BLI
SA n
SA
n+1
BLI n+1 = VDD
SA nodes at
LV(VDDSA/VDD/VSS
BLn = VH
BL = VH
BLn+1
CELSRC
CELSRC
CELSRC
p+
n+
n+
n+
CPWELL=Vera

FIG. 30

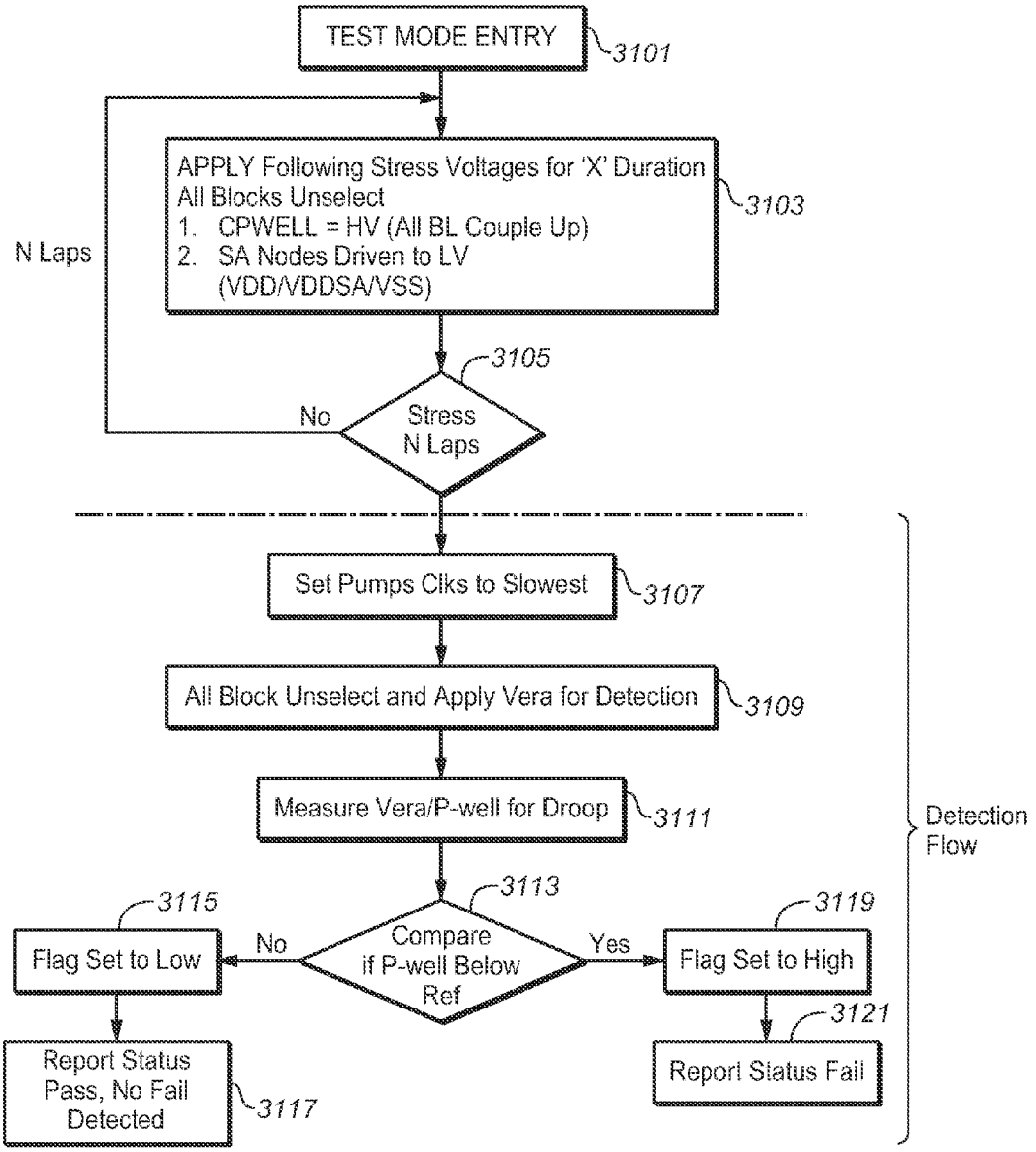

TEST MODE ENTRY
3101
APPLY Following Stress Voltages for 'X' Duration
All Blocks Unselect
1. CPWELL = HV (All BL Couple Up)
2. SA Nodes Driven to LV
(VDD/VDDSA/VSS)
3103
N Laps
3105
No
Stress
N Laps
Set Pumps Clks to Slowest
3107
All Block Unselect and Apply Vera for Detection
3109
Measure Vera/P-well for Droop
3111
3113
Compare
if P-well Below
Ref
No
Yes
3115
Flag Set to Low
3119
Flag Set to High
3121
Report Status Fail
Report Status
Pass, No Fail
Detected
3117
Detection
Flow

FIG. 31

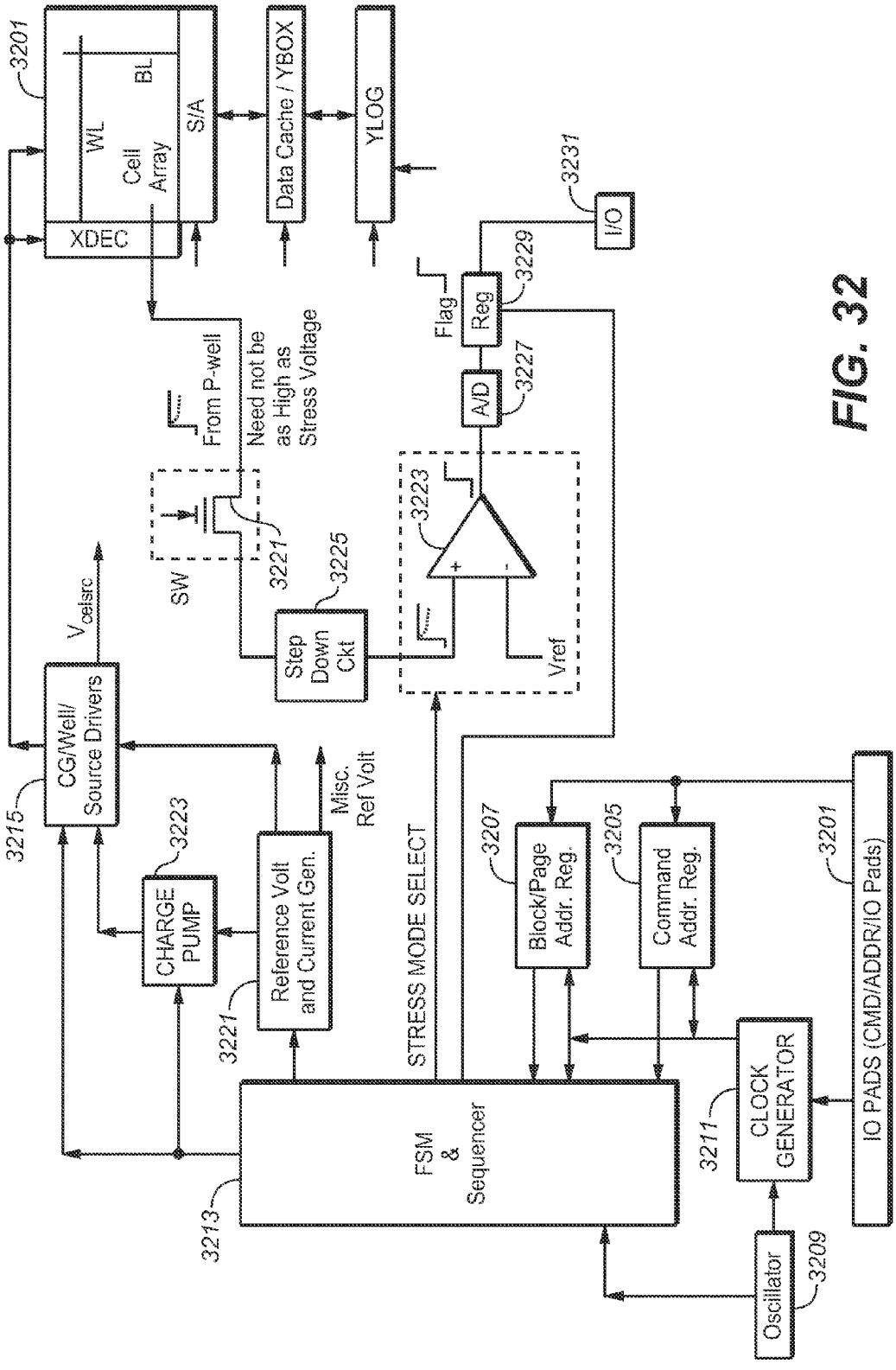

3201
WL
BL
Cell Array
S/A
XDEC
Data Cache / YBOX
YLOG
3231
I/O
3229
Reg
Flag
3227
A/D
3223
+
-
Vref
From P-well
Need not be as High as Stress Voltage
SW
3221
3225
Step Down Ckt
V_{celsrc}
CG/Well/ Source Drivers
3215
CHARGE PUMP
3223
Reference Volt and Current Gen.
3221
Misc. Ref Volt
STRESS MODE SELECT
3207
Block/Page Addr. Reg.
3205
Command Addr. Reg.
3201
IO PADS (CMD/ADDR/IO Pads)
FSM & Sequencer
3213
CLOCK GENERATOR
3211
Oscillator
3209

FIG. 32

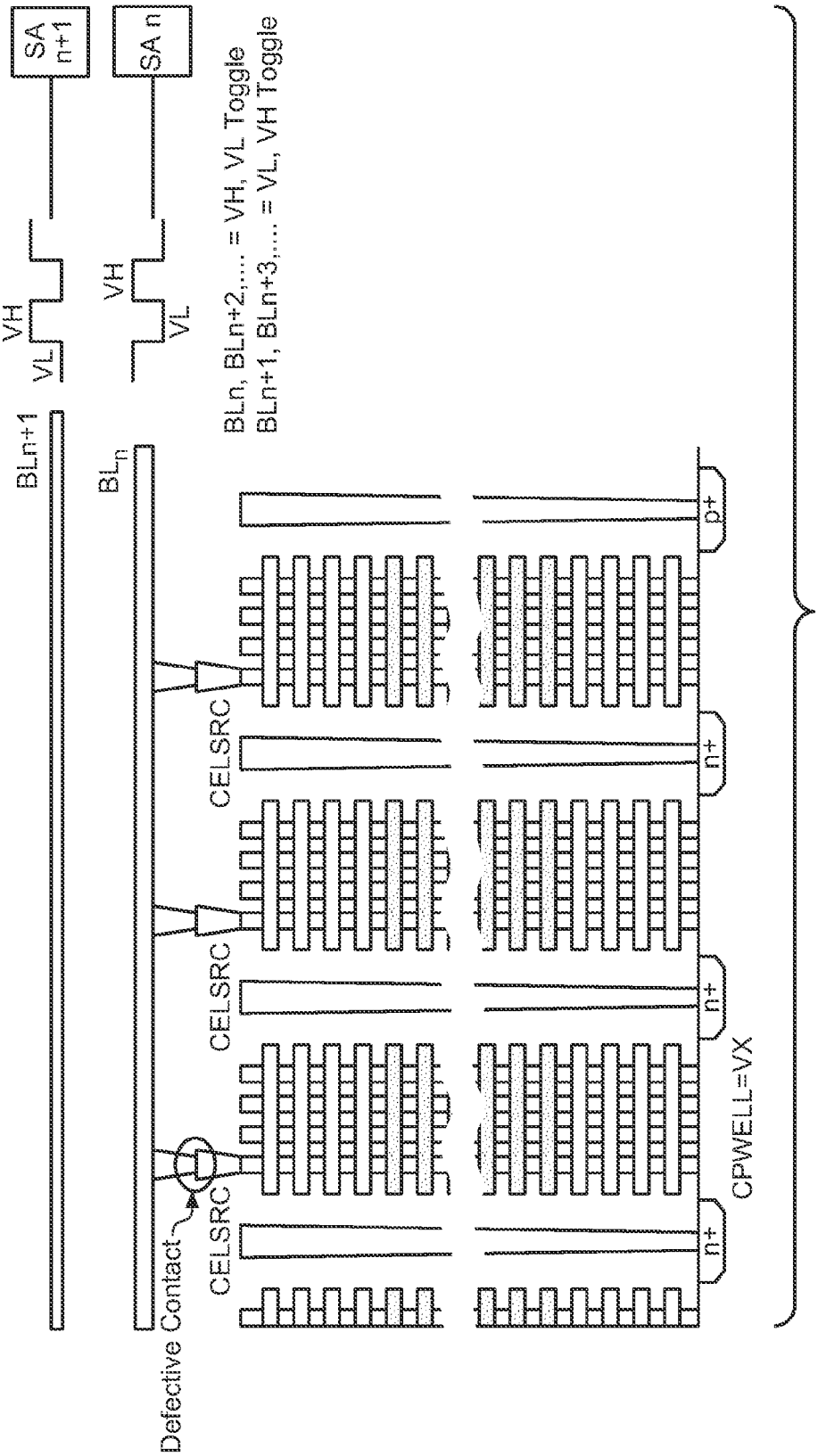

SA n+1
SA n
VH
VL
VH
VL
BLn+1
BLn
BLn, BLn+2,.... = VH, VL Toggle
BLn+1, BLn+3,.... = VL, VH Toggle
CELSRC
CELSRC
CELSRC
Defective Contact
n+
n+
n+
p+
CPWELL=VX

FIG. 33

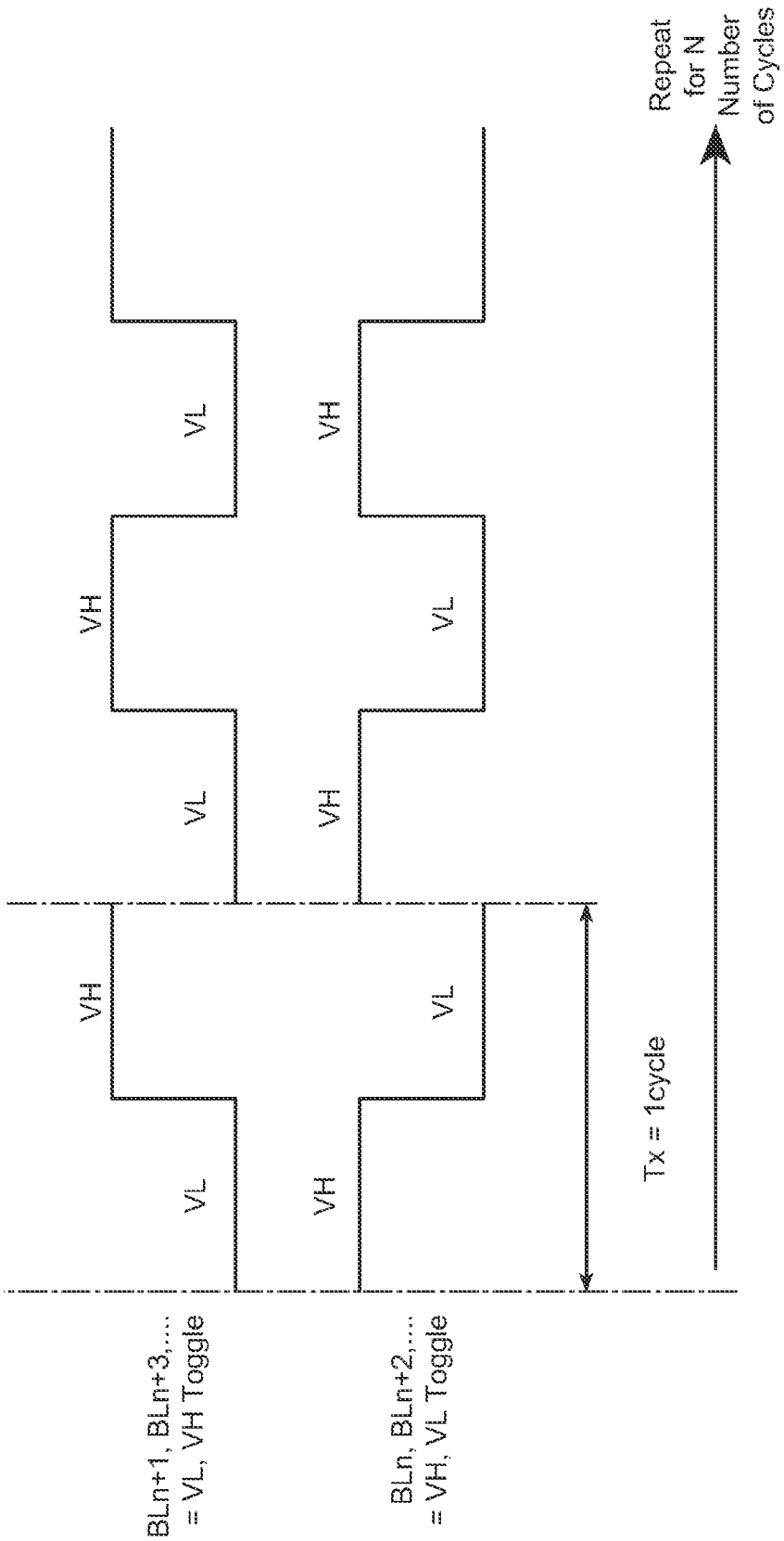

Repeat for N Number of Cycles
VH
VL
VH
VL
VH
VL
VH
VL
VH
VL
Tx = 1cycle
BLn+1, BLn+3,.... = VL, VH Toggle
BLn, BLn+2,.... = VH, VL Toggle

FIG. 34

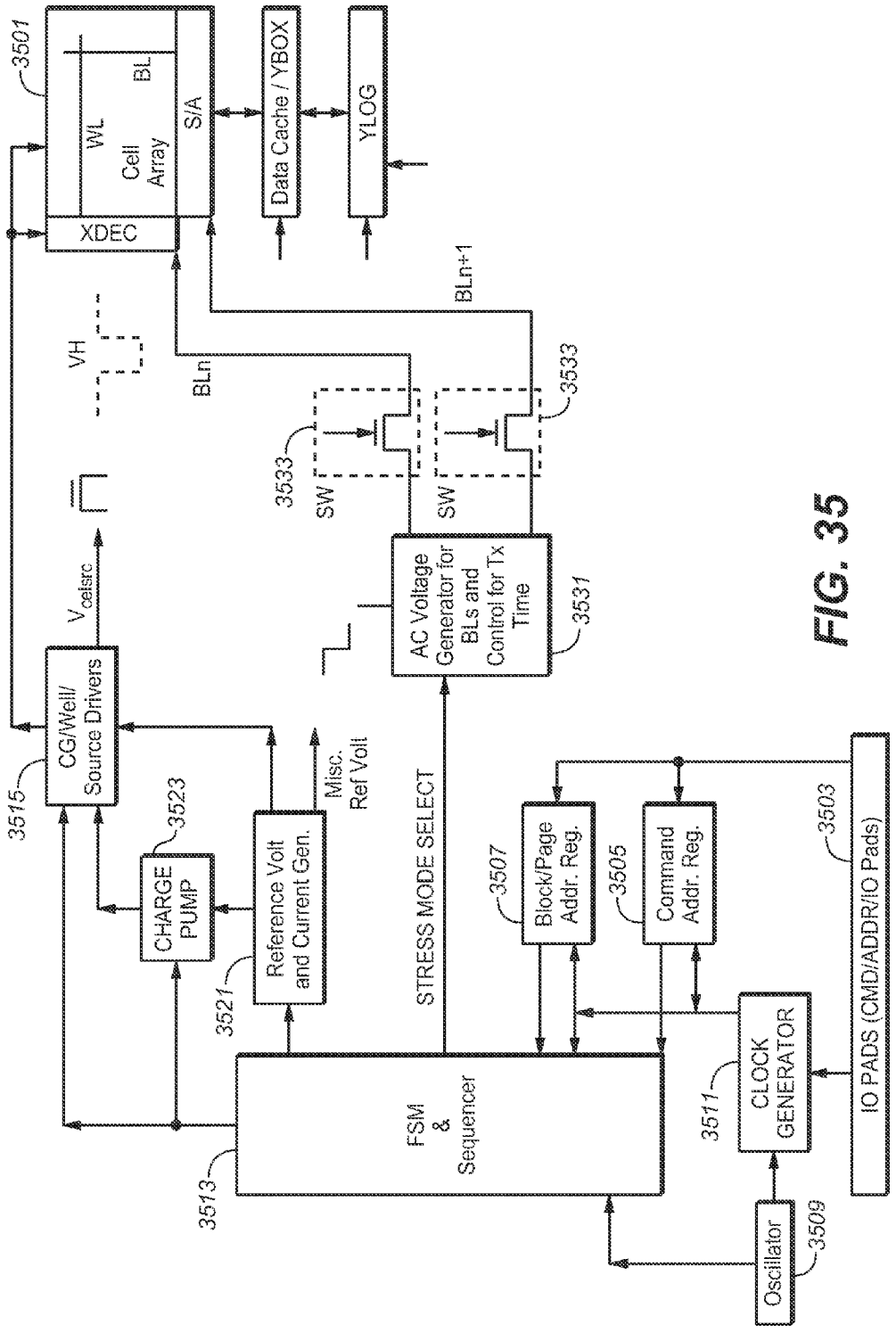

3501
WL
BL
Cell Array
S/A
XDEC
Data Cache / YBOX
YLOG
VH
BLn
BLn+1
3533
SW
SW
3533
AC Voltage Generator for BLs and Control for Tx Time
3531
V_{oelsrc}
CG/Well/ Source Drivers
3515
CHARGE PUMP
3523
Reference Volt and Current Gen.
3521
Misc. Ref Volt
STRESS MODE SELECT
3507
Block/Page Addr. Reg.
3505
Command Addr. Reg.
3503
IO PADS (CMD/ADDR/IO Pads)
FSM & Sequencer
3513
CLOCK GENERATOR
3511
Oscillator
3509

FIG. 35

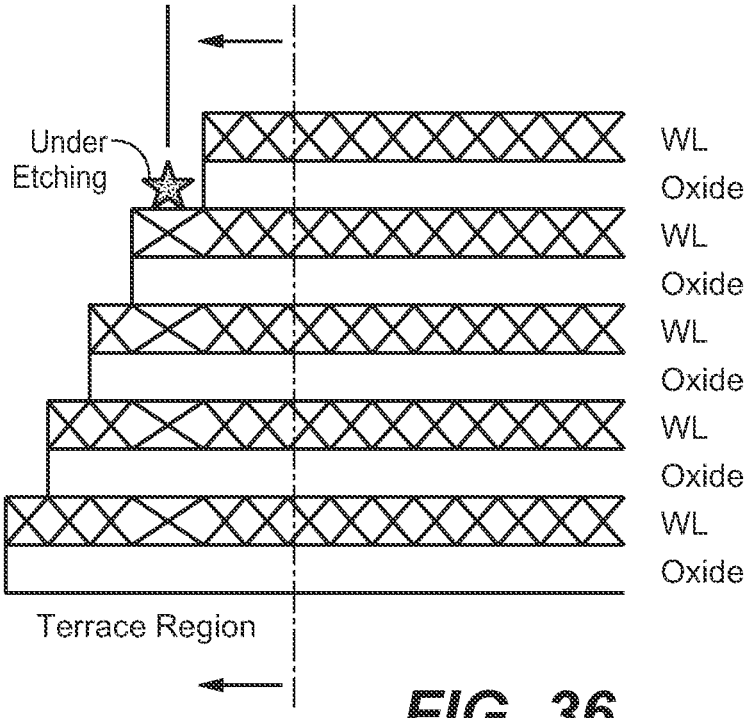

Under
Etching
WL
Oxide
WL
Oxide
WL
Oxide
WL
Oxide
WL
Oxide
Terrace Region

FIG. 36

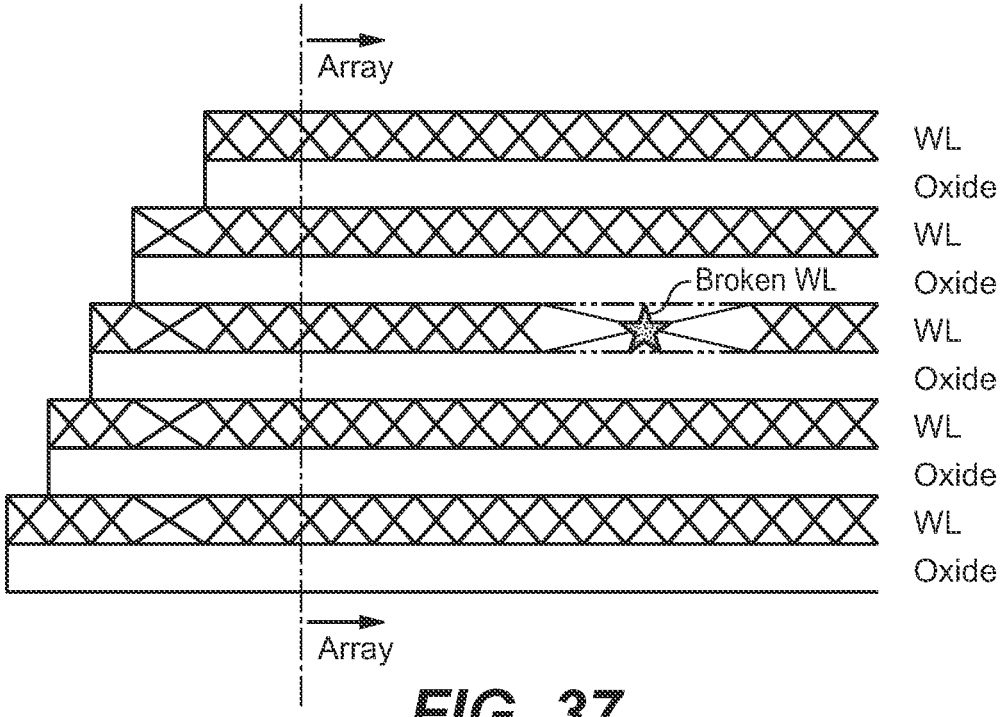

Array
WL
Oxide
WL
Broken WL
Oxide
WL
Oxide
WL
Oxide
WL
Oxide
Array

FIG. 37

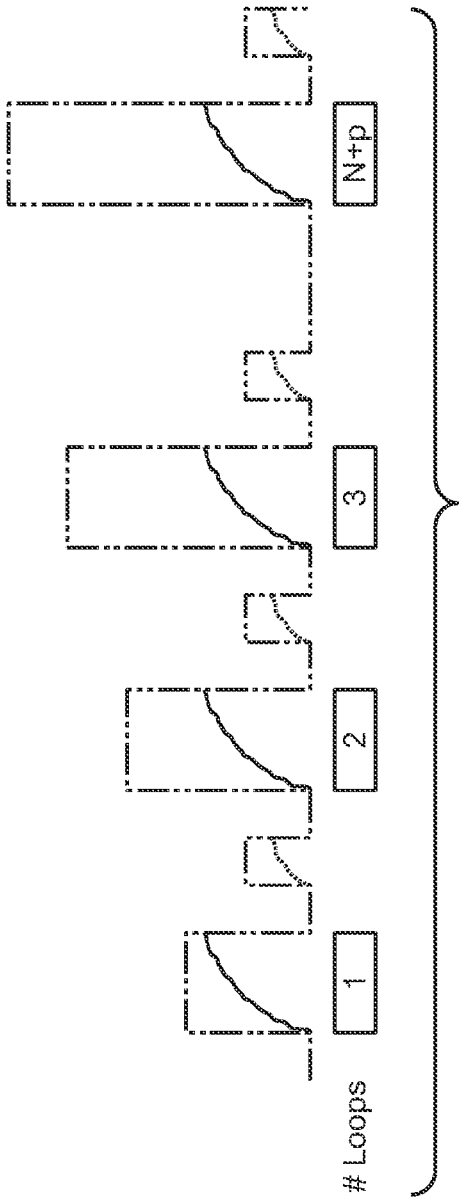


FIG. 38

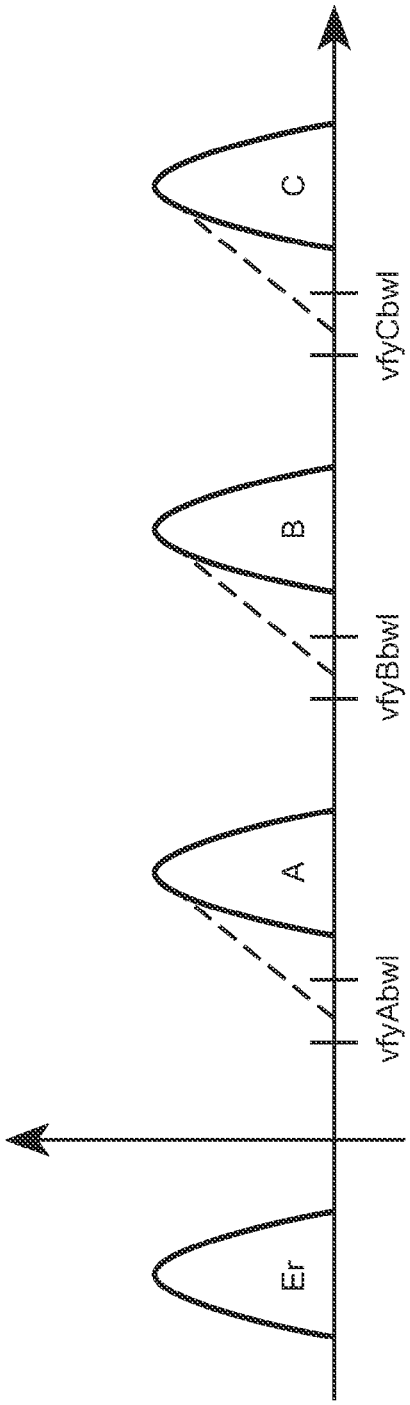


FIG. 39

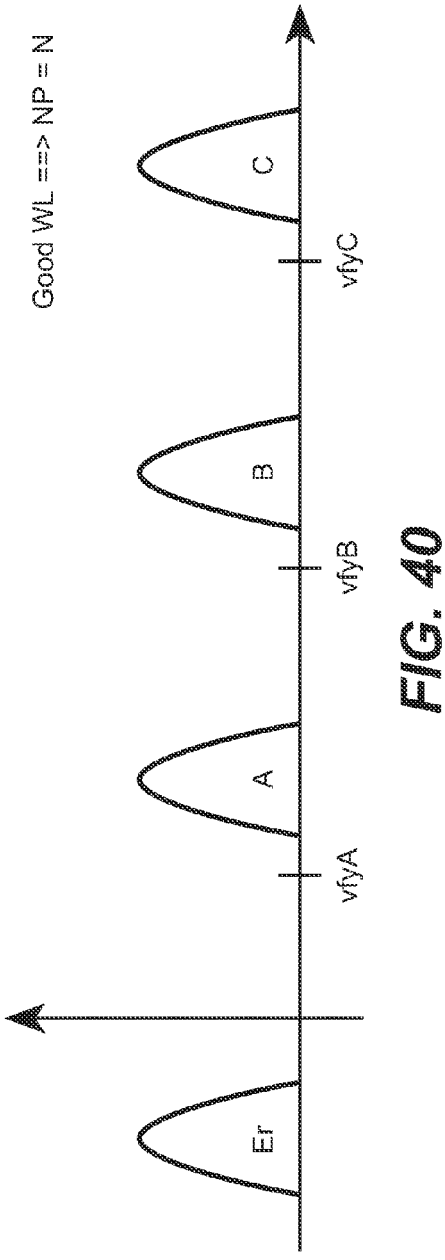

Good WL ==> NP = N
Er
A
B
C
vfyA
vfyB
vfyC

FIG. 40

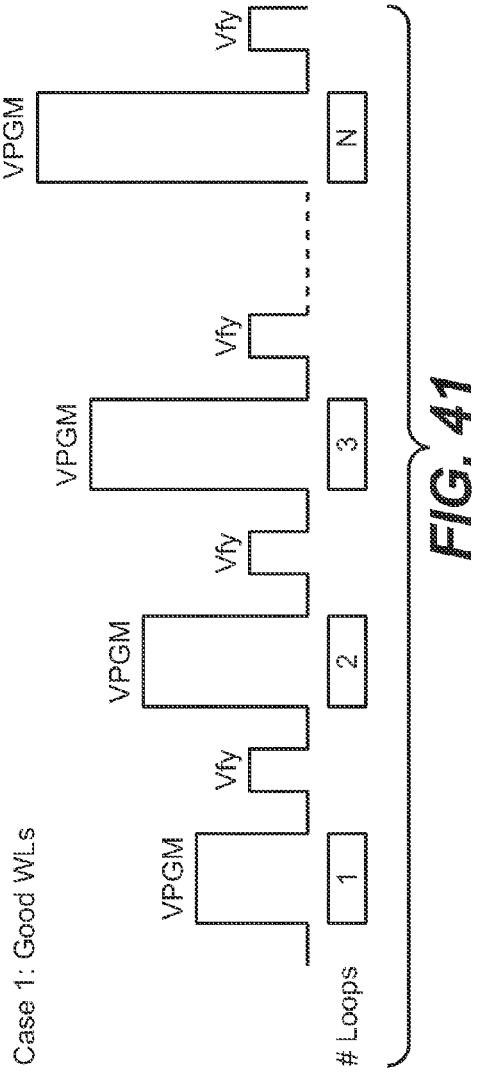

Case 1: Good WLs
Loops
1
2
3
N
VPGM
Vfy

FIG. 41

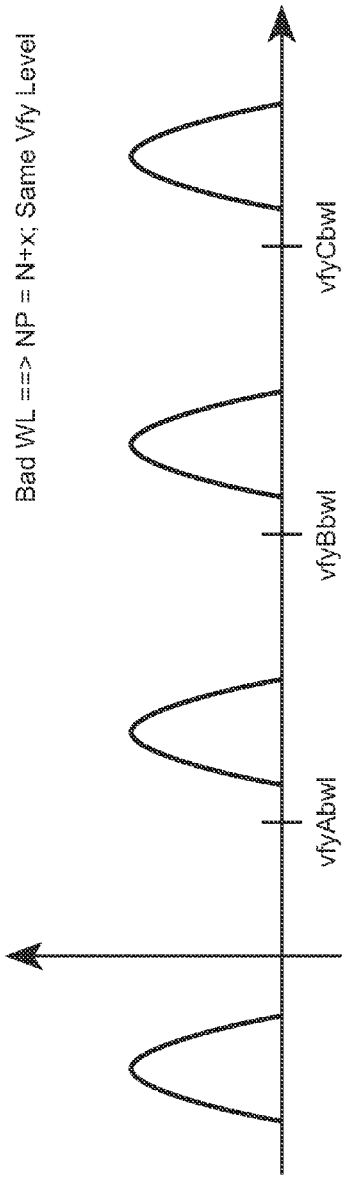

Bad WL ==> NP = N+x; Same Vfy Level
vfyAbwl
vfyBbwl
vfyCbwl

FIG. 42

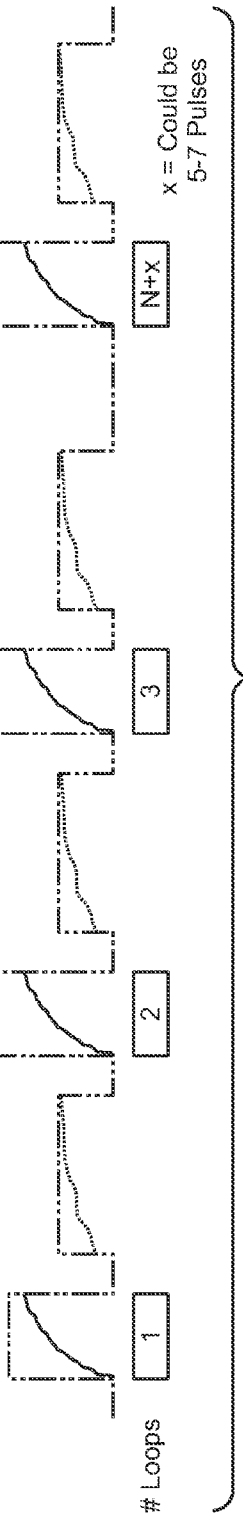

Loops
1
2
3
N+x
x = Could be 5-7 Pulses

FIG. 43

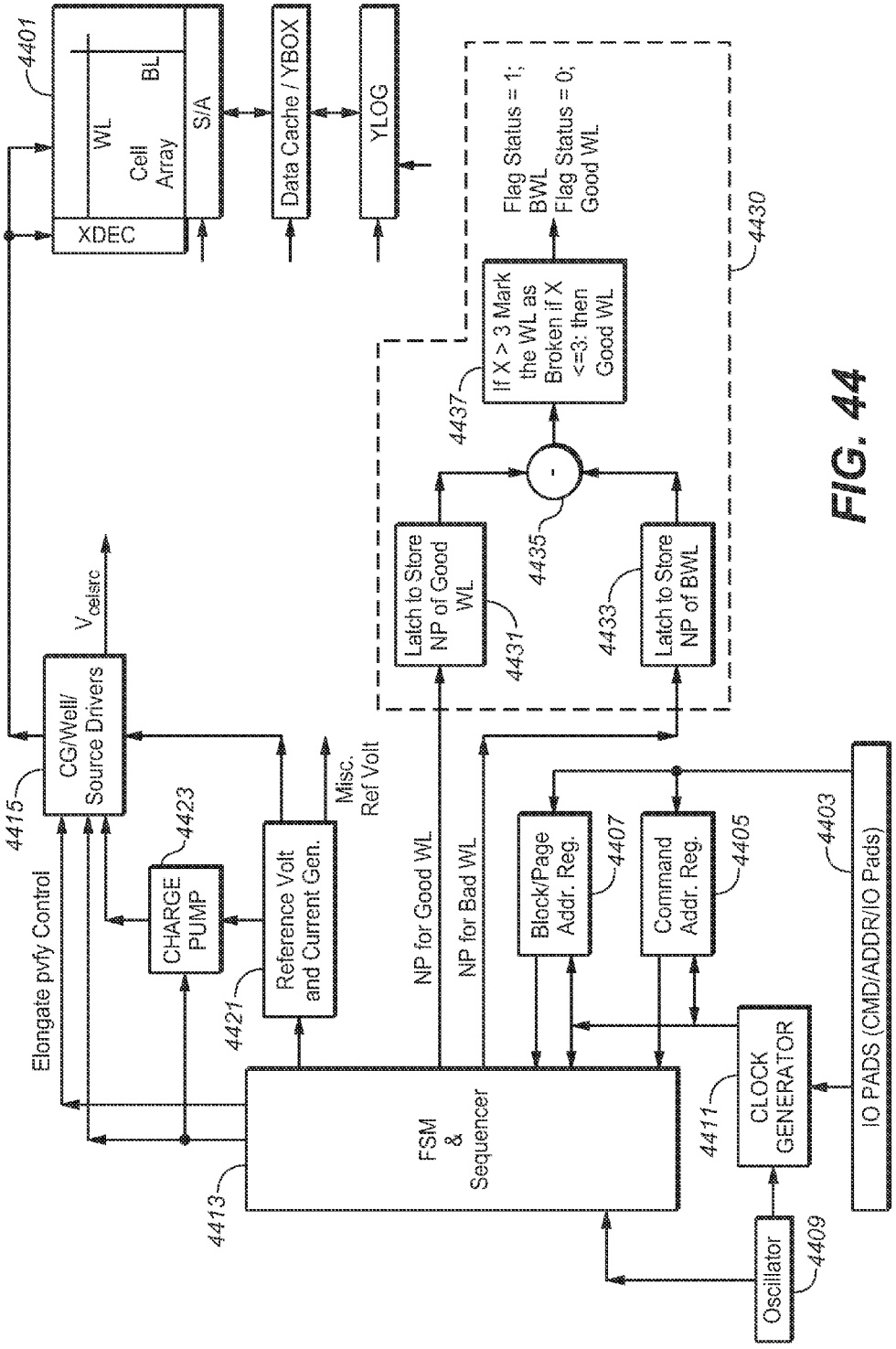

4401
XDEC
WL
Cell Array
BL
S/A
Data Cache / YBOX
YLOG
4430
Flag Status = 1; BWL
Flag Status = 0; Good WL
4437
If X > 3 Mark the WL as Broken if X <=3: then Good WL
4435
4431
Latch to Store NP of Good WL
4433
Latch to Store NP of BWL
V_{celsrc}
4415
CG/Well/ Source Drivers
Misc. Ref Volt
4423
CHARGE PUMP
4421
Reference Volt and Current Gen.
Elongate pvfy Control
NP for Good WL
NP for Bad WL
4407
Block/Page Addr. Reg.
4405
Command Addr. Reg.
4403
IO PADS (CMD/ADDR/IO Pads)
4411
CLOCK GENERATOR
4413
FSM & Sequencer
4409
Oscillator

FIG. 44

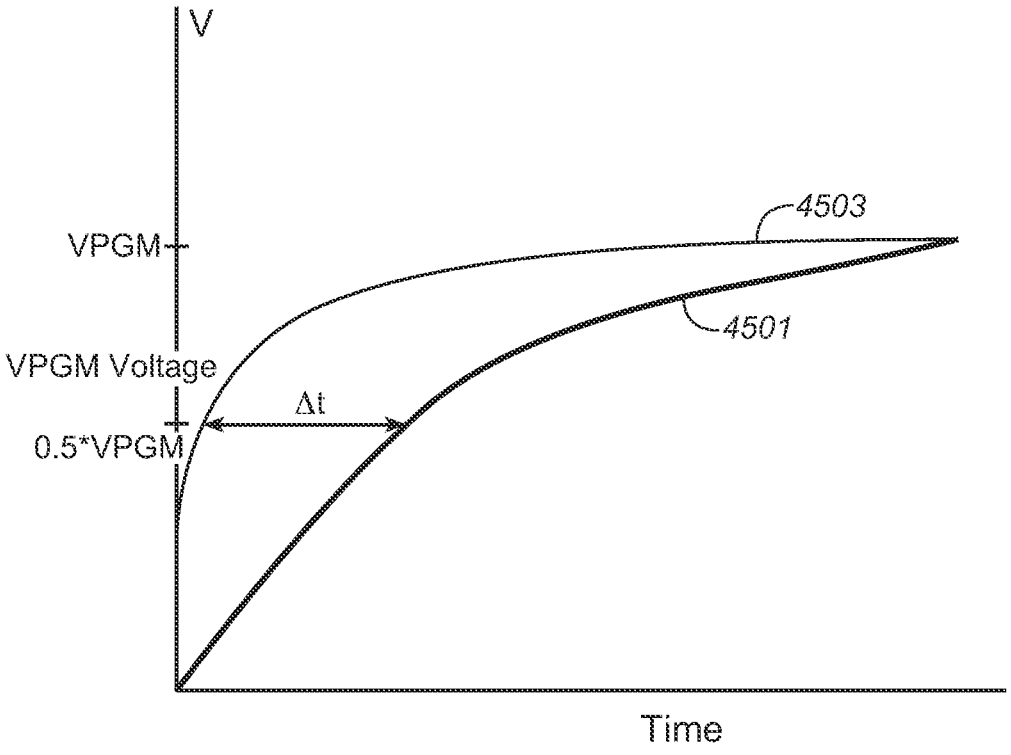


FIG. 45

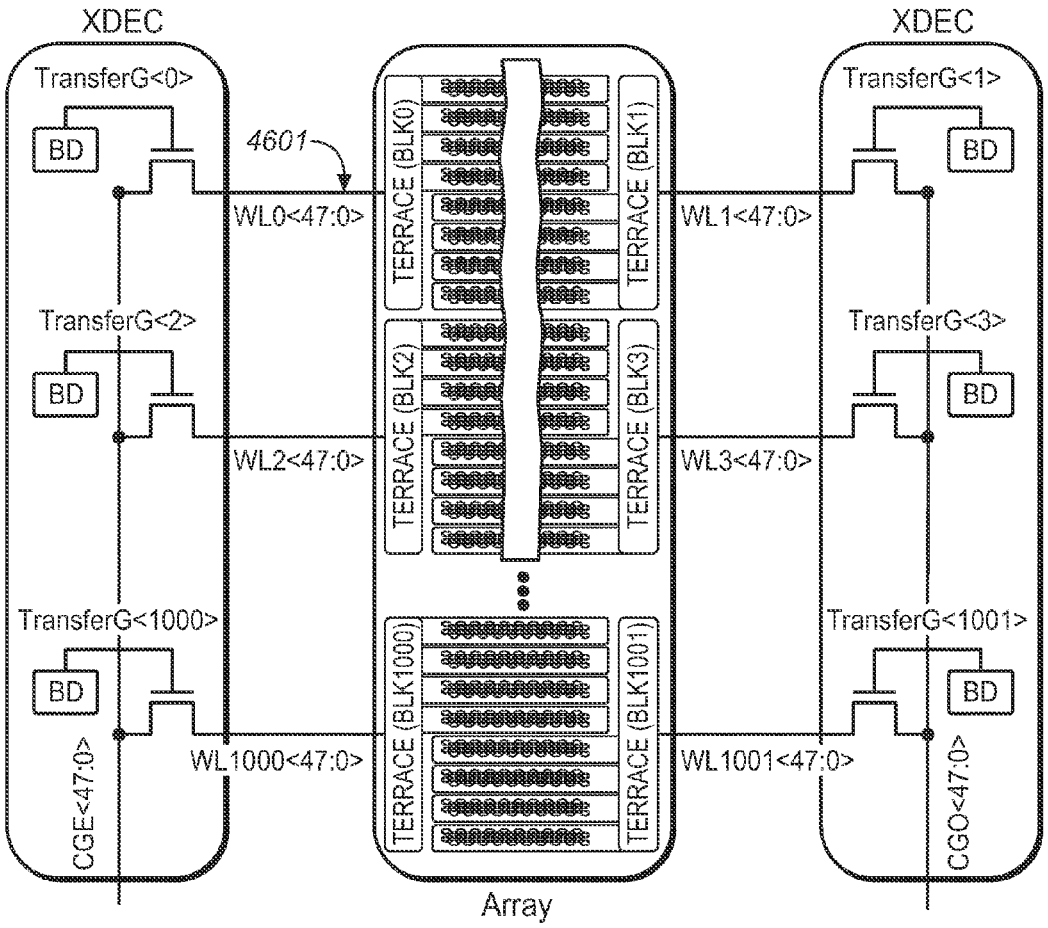

XDEC
XDEC
TransferG<0>
TransferG<1>
BD
BD
4601
WL0<47:0>
WL1<47:0>
TERRACE (BLK0)
TERRACE (BLK1)
TransferG<2>
TransferG<3>
BD
BD
WL2<47:0>
WL3<47:0>
TERRACE (BLK2)
TERRACE (BLK3)
TransferG<1000>
TransferG<1001>
BD
BD
WL1000<47:0>
WL1001<47:0>
TERRACE (BLK1000)
TERRACE (BLK1001)
CGE<47:0>
CGO<47:0>
Array

FIG. 46

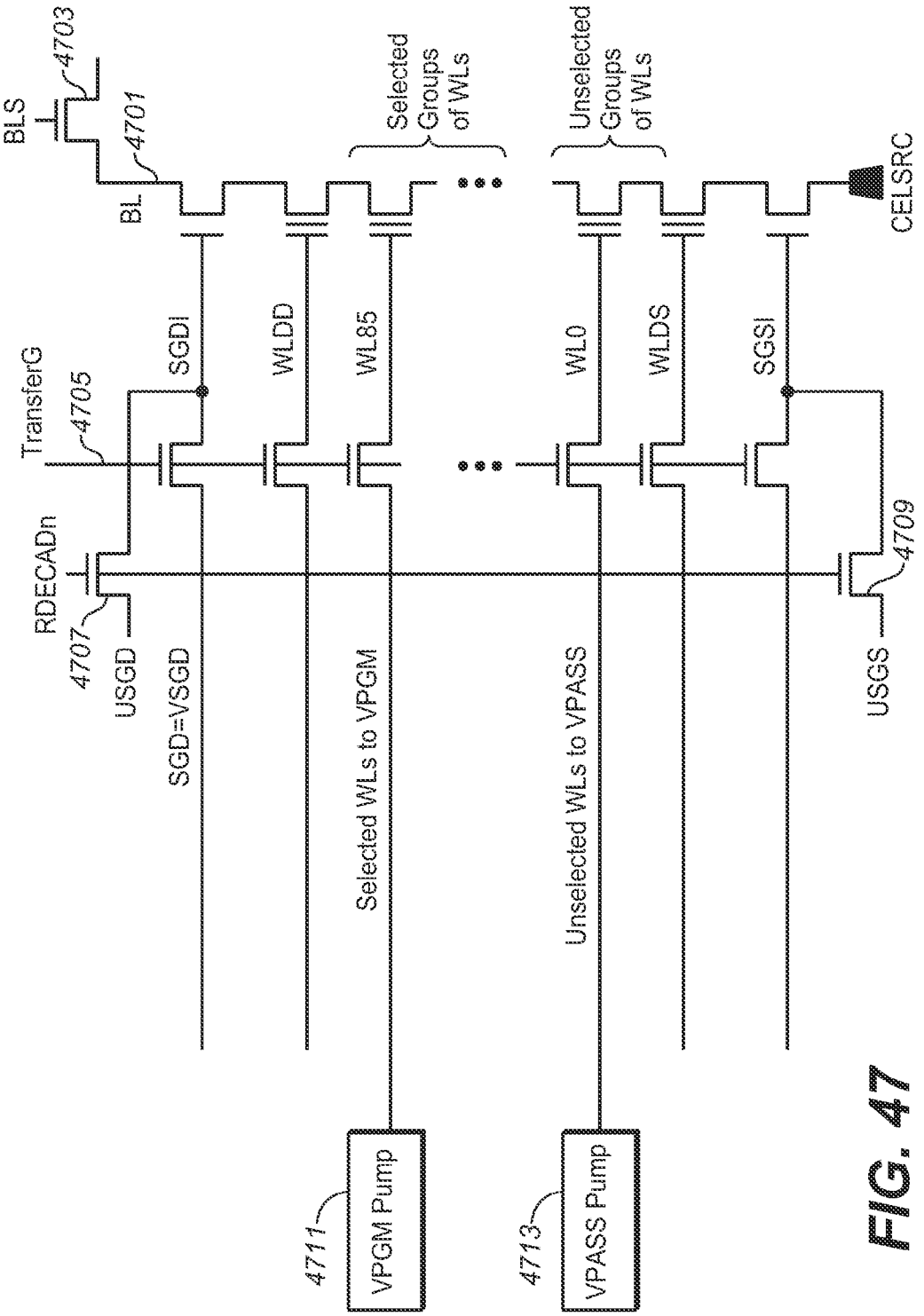

BLS
4703
4701
BL
Selected Groups of WLs
Unselected Groups of WLs
CELSRC
TransferG
4705
SGDI
WLDD
WL85
WL0
WLDS
SGSI
RDECADn
4707
USGD
SGD=VSGD
Selected WLs to VPGM
Unselected WLs to VPASS
4709
USGS
4711
VPGM Pump
4713
VPASS Pump

FIG. 47

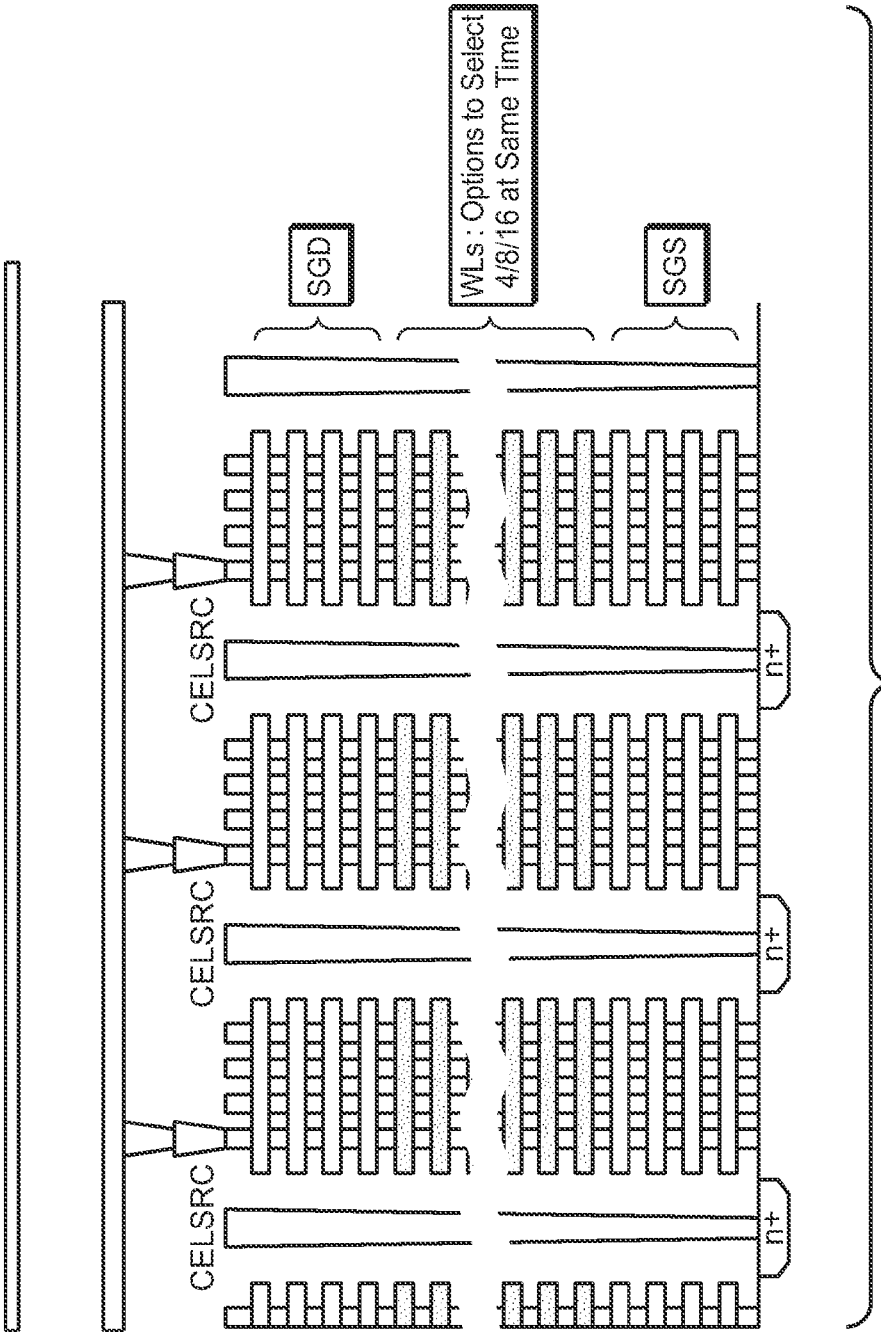

SGD
WLs : Options to Select 4/8/16 at Same Time
SGS
CELSRC
CELSRC
CELSRC
n+
n+
n+

FIG. 48

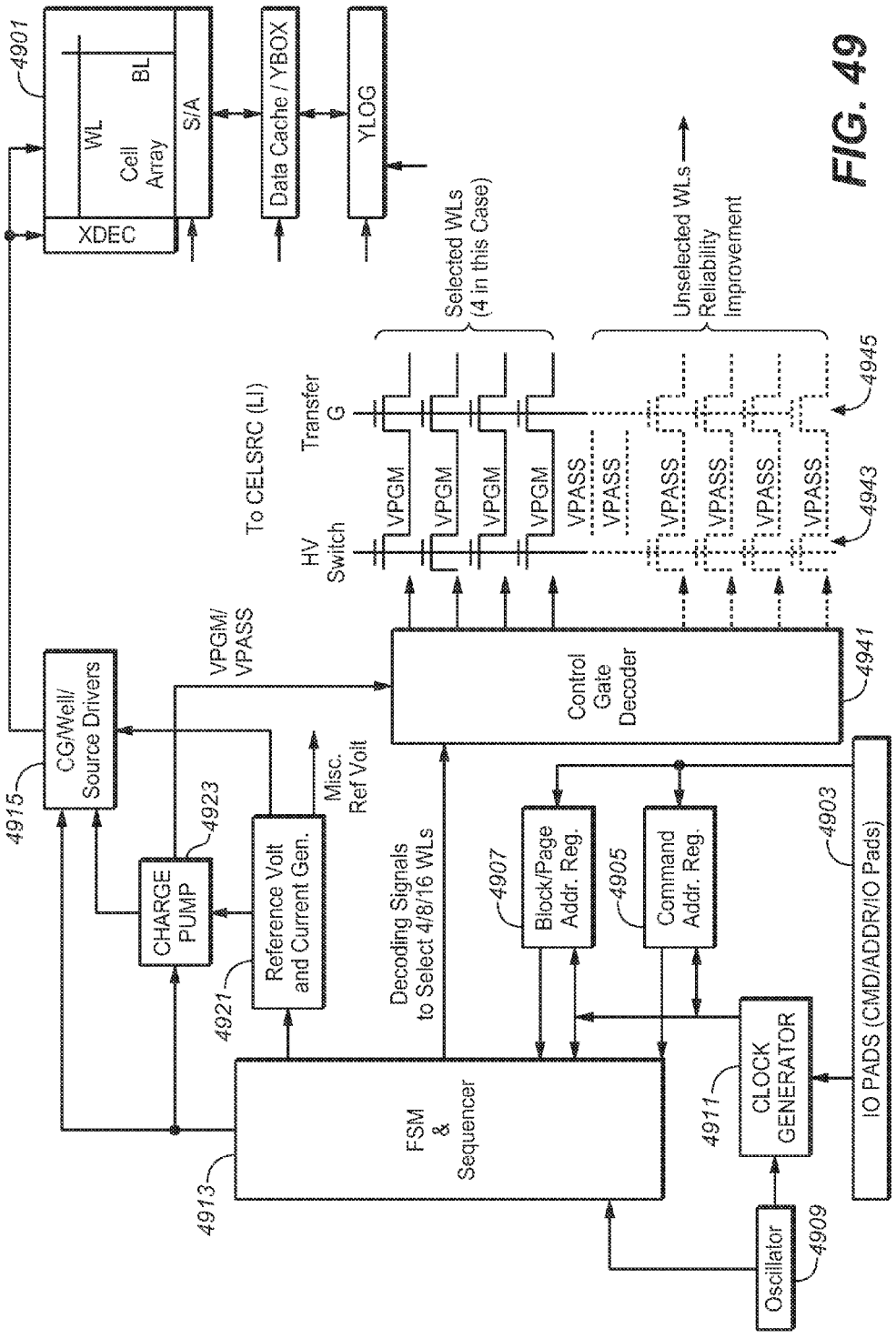

4901
WL
Cell Array
BL
S/A
XDEC
Data Cache / YBOX
YLOG
Selected WLs (4 in this Case)
Unselected WLs Reliability Improvement
To CELSRC (LI)
Transfer G
HV Switch
VPGM
VPASS
4945
4943
VPGM/ VPASS
Control Gate Decoder
4941
4915
CG/Well/ Source Drivers
Misc. Ref Volt
4923
CHARGE PUMP
Reference Volt and Current Gen.
4921
Decoding Signals to Select 4/8/16 WLs
4907
Block/Page Addr. Reg.
4905
Command Addr. Reg.
4903
IO PADS (CMD/ADDR/IO Pads)
FSM & Sequencer
4913
4911
CLOCK GENERATOR
Oscillator
4909

FIG. 49

NON-VOLATILE MEMORY WITH MULTI-WORD LINE SELECT FOR DEFECT DETECTION OPERATIONS

BACKGROUND

This application relates to the operation of re-programmable non-volatile memory systems such as semiconductor flash memory that record data using charge stored in charge storage elements of memory cells.

Solid-state memory capable of nonvolatile storage of charge, particularly in the form of EEPROM and flash EEPROM packaged as a small form factor card, has recently become the storage of choice in a variety of mobile and handheld devices, notably information appliances and consumer electronics products. Unlike RAM (random access memory) that is also solid-state memory, flash memory is non-volatile, and retains its stored data even after power is turned off. Also, unlike ROM (read only memory), flash memory is rewritable similar to a disk storage device. In spite of the higher cost, flash memory is increasingly being used in mass storage applications.

Flash EEPROM is similar to EEPROM (electrically erasable and programmable read-only memory) in that it is a non-volatile memory that can be erased and have new data written or "programmed" into their memory cells. Both utilize a floating (unconnected) conductive gate, in a field effect transistor structure, positioned over a channel region in a semiconductor substrate, between source and drain regions. A control gate is then provided over the floating gate. The threshold voltage characteristic of the transistor is controlled by the amount of charge that is retained on the floating gate. That is, for a given level of charge on the floating gate, there is a corresponding voltage (threshold) that must be applied to the control gate before the transistor is turned "on" to permit conduction between its source and drain regions. Flash memory such as Flash EEPROM allows entire blocks of memory cells to be erased at the same time.

The floating gate can hold a range of charges and therefore can be programmed to any threshold voltage level within a threshold voltage window. The size of the threshold voltage window is delimited by the minimum and maximum threshold levels of the device, which in turn correspond to the range of the charges that can be programmed onto the floating gate. The threshold window generally depends on the memory device's characteristics, operating conditions and history. Each distinct, resolvable threshold voltage level range within the window may, in principle, be used to designate a definite memory state of the cell.

In order to improve read and program performance, multiple charge storage elements or memory transistors in an array are read or programmed in parallel. Thus, a "page" of memory elements are read or programmed together. In existing memory architectures, a row typically contains several interleaved pages or it may constitute one page. All memory elements of a page are read or programmed together.

Nonvolatile memory devices are also manufactured from memory cells with a dielectric layer for storing charge. Instead of the conductive floating gate elements described earlier, a dielectric layer is used. An ONO dielectric layer extends across the channel between source and drain diffusions. The charge for one data bit is localized in the dielectric layer adjacent to the drain, and the charge for the other data bit is localized in the dielectric layer adjacent to the source. For example, a nonvolatile memory cell may have a trapping dielectric sandwiched between two silicon dioxide layers. Multi-state data storage is implemented by separately reading the binary states of the spatially separated charge storage regions within the dielectric.

SUMMARY

A first set of aspects relate to a non-volatile memory circuit having an array of non-volatile memory cells formed according to a NAND architecture as a plurality of blocks, each block formed of a plurality of NAND strings having multiple memory cells connected in series and connected along word lines. Bias circuitry provides bias voltage levels for use in the operation of the array, where the bias circuit is connectable by decoding circuitry to the array to selectively apply the bias voltage levels to the array. During a programming operation the decoding circuitry applies a programming voltage to a selected word line of a selected block while applying a pass voltage to non-selected word lines of the selected block, the programming voltage being higher that the pass voltage. During a stress operation the decoding circuitry applies a stress voltage concurrently to a first plurality of selected word lines of a selected block while applying the pass voltage to one or more non-selected word lines of the selected block, the stress voltage being higher than the pass voltage.

Various aspects, advantages, features and embodiments are included in the following description of exemplary examples thereof, which description should be taken in conjunction with the accompanying drawings. All patents, patent applications, articles, other publications, documents and things referenced herein are hereby incorporated herein by this reference in their entirety for all purposes. To the extent of any inconsistency or conflict in the definition or use of terms between any of the incorporated publications, documents or things and the present application, those of the present application shall prevail.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** illustrates schematically the main hardware components of a memory system suitable for implementing various aspects described in the following.

FIG. **2** illustrates schematically a non-volatile memory cell.

FIG. **3** illustrates the relation between the source-drain current I_D and the control gate voltage V_{CG} for four different charges Q**1**-Q**4** that the floating gate may be selectively storing at any one time at fixed drain voltage.

FIG. **4** illustrates schematically a string of memory cells organized into a NAND string.

FIG. **5** illustrates an example of a NAND array **210** of memory cells, constituted from NAND strings **50** such as that shown in FIG. **4**.

FIG. **6** illustrates a page of memory cells, organized in the NAND configuration, being sensed or programmed in parallel.

FIGS. **7**A-**7**C illustrate an example of programming a population of memory cells.

FIG. **8** shows an example of a physical structure of a 3-D NAND string.

FIGS. **9**-**12** look at a particular monolithic three dimensional (3D) memory array of the NAND type (more specifically of the "BiCS" type).

FIG. **13** is a side view of a block, similar to FIG. **11**, but with some of the features highlighted.

FIG. **14** shows the toggling of voltage levels being applied to the two sets of word lines to apply an AC stress.

FIGS. **15** and **16** show an exemplary flow for an on-chip AC stress and defect determination process for word line to word line shorts within a block.

FIG. **17** is a schematic representation of some of the elements on the memory chip that are involved in the process of FIGS. **15** and **16**.

FIG. **18** is a top level, top down diagram of how blocks are paired and placed next to each other in the exemplary embodiment of FIGS. **9-12**.

FIGS. **19** and **20** schematically represents two example of the word line levels as applied to even and odd block in an inter-block stress mode,

FIGS. **21**A and **21**B are a schematic representation of some of the elements on the memory chip involved in an inter-block word line short determination process.

FIG. **22** is an oblique view of a simplified version of one block with four fingers, where each NAND string only has two memory cells and only a single select gate at either end.

FIGS. **23** and **24** illustrate the consequences of a short between two word lines of the same block and a short between select gates.

FIG. **25** shows a side view of the exemplary embodiment with an overview of the applied voltages to stress for local interconnect to word line physical shorts.

FIGS. **26** and **27** are an exemplary flow of a test mode to screen local interconnect to word line physical shorts.

FIG. **28** is a schematic representation of some of the elements on the memory chip that are involved in the process of FIGS. **26** and **27**.

FIG. **29** is a schematic representation of a possible bit line to low voltage short.

FIG. **30** is similar to FIG. **29**, but marked with some of the voltage levels involved in the stress phase of determining bit line shorts.

FIG. **31** is an exemplary flow for test process for determining bit line to low voltage shorts.

FIG. **32** is a schematic representation of some of the elements on the memory chip involved in the process of FIG. **31**B.

FIG. **33** shows a view of a 3D array structure with an example of a bit line-memory string contact highlighted.

FIG. **34** represents an exemplary set of waveforms for the AC voltage applied to the bi lines.

FIG. **35** is a schematic representation of some of the elements on the memory chip for applying an AC stress to bit lines.

FIGS. **36** and **37** illustrate some examples of broken word lines in a memory array.

FIGS. **38** and **39** represent a program/verify waveform, how a broken word line responds, and effect on the distribution of programmed cells.

FIGS. **40** and **41** show the desired distribution of states and a standard program/verify waveform.

FIGS. **42** and **43** show a distribution of states and a modified test program/verify waveform with an elongated verify.

FIG. **44** is a schematic representation of an embodiment for some of the circuitry that can be used in determining whether a word or select line is defective based on the test program.

FIG. **45** illustrates the rate at which a broken word line may ramp up relative to a good word line.

FIG. **46** shows an example of a point after the pass transistor where the slopes in FIG. **45** can be measured for a word line or select gate line.

FIGS. **47** and **48** respectively illustrate 2D NAND and BiCS based examples.

FIG. **49** is a block diagram that highlights some circuitry involved in a multi-word line select test mode.

DETAILED DESCRIPTION

Memory System

FIG. **1** illustrates schematically the main hardware components of a memory system suitable for implementing the following. The memory system **90** typically operates with a host **80** through a host interface. The memory system may be in the form of a removable memory such as a memory card, or may be in the form of an embedded memory system. The memory system **90** includes a memory **102** whose operations are controlled by a controller **100**. The memory **102** comprises one or more array of non-volatile memory cells distributed over one or more integrated circuit chip. The controller **100** may include interface circuits **110**, a processor **120**, ROM (read-only-memory) **122**, RAM (random access memory) **130**, programmable nonvolatile memory **124**, and additional components. The controller is typically formed as an ASIC (application specific integrated circuit) and the components included in such an ASIC generally depend on the particular application.

With respect to the memory section **102**, semiconductor memory devices include volatile memory devices, such as dynamic random access memory ("DRAM") or static random access memory ("SRAM") devices, non-volatile memory devices, such as resistive random access memory ("ReRAM"), electrically erasable programmable read only memory ("EEPROM"), flash memory (which can also be considered a subset of EEPROM), ferroelectric random access memory ("FRAM"), and magnetoresistive random access memory ("MRAM"), and other semiconductor elements capable of storing information. Each type of memory device may have different configurations. For example, flash memory devices may be configured in a NAND or a NOR configuration.

The memory devices can be formed from passive and/or active elements, in any combinations. By way of non-limiting example, passive semiconductor memory elements include ReRAM device elements, which in some embodiments include a resistivity switching storage element, such as an anti-fuse, phase change material, etc., and optionally a steering element, such as a diode, etc. Further by way of non-limiting example, active semiconductor memory elements include EEPROM and flash memory device elements, which in some embodiments include elements containing a charge storage region, such as a floating gate, conductive nanoparticles, or a charge storage dielectric material.

Multiple memory elements may be configured so that they are connected in series or so that each element is individually accessible. By way of non-limiting example, flash memory devices in a NAND configuration (NAND memory) typically contain memory elements connected in series. A NAND memory array may be configured so that the array is composed of multiple strings of memory in which a string is composed of multiple memory elements sharing a single bit line and accessed as a group. Alternatively, memory elements may be configured so that each element is individually accessible, e.g., a NOR memory array. NAND and NOR memory configurations are exemplary, and memory elements may be otherwise configured.

The semiconductor memory elements located within and/or over a substrate may be arranged in two or three dimensions, such as a two dimensional memory structure or a three dimensional memory structure.

In a two dimensional memory structure, the semiconductor memory elements are arranged in a single plane or a single memory device level. Typically, in a two dimensional memory structure, memory elements are arranged in a plane (e.g., in an x-z direction plane) which extends substantially parallel to a major surface of a substrate that supports the memory elements. The substrate may be a wafer over or in which the layer of the memory elements are formed or it may be a carrier substrate which is attached to the memory elements after they are formed. As a non-limiting example, the substrate may include a semiconductor such as silicon.

The memory elements may be arranged in the single memory device level in an ordered array, such as in a plurality of rows and/or columns. However, the memory elements may be arrayed in non-regular or non-orthogonal configurations. The memory elements may each have two or more electrodes or contact lines, such as bit lines and word lines.

A three dimensional memory array is arranged so that memory elements occupy multiple planes or multiple memory device levels, thereby forming a structure in three dimensions (i.e., in the x, y and z directions, where the y direction is substantially perpendicular and the x and z directions are substantially parallel to the major surface of the substrate).

As a non-limiting example, a three dimensional memory structure may be vertically arranged as a stack of multiple two dimensional memory device levels. As another non-limiting example, a three dimensional memory array may be arranged as multiple vertical columns (e.g., columns extending substantially perpendicular to the major surface of the substrate, i.e., in the y direction) with each column having multiple memory elements in each column. The columns may be arranged in a two dimensional configuration, e.g., in an x-z plane, resulting in a three dimensional arrangement of memory elements with elements on multiple vertically stacked memory planes. Other configurations of memory elements in three dimensions can also constitute a three dimensional memory array.

By way of non-limiting example, in a three dimensional NAND memory array, the memory elements may be coupled together to form a NAND string within a single horizontal (e.g., x-z) memory device levels. Alternatively, the memory elements may be coupled together to form a vertical NAND string that traverses across multiple horizontal memory device levels. Other three dimensional configurations can be envisioned wherein some NAND strings contain memory elements in a single memory level while other strings contain memory elements which span through multiple memory levels. Three dimensional memory arrays may also be designed in a NOR configuration and in a ReRAM configuration.

Typically, in a monolithic three dimensional memory array, one or more memory device levels are formed above a single substrate. Optionally, the monolithic three dimensional memory array may also have one or more memory layers at least partially within the single substrate. As a non-limiting example, the substrate may include a semiconductor such as silicon. In a monolithic three dimensional array, the layers constituting each memory device level of the array are typically formed on the layers of the underlying memory device levels of the array. However, layers of adjacent memory device levels of a monolithic three dimensional memory array may be shared or have intervening layers between memory device levels.

Then again, two dimensional arrays may be formed separately and then packaged together to form a non-monolithic memory device having multiple layers of memory. For example, non-monolithic stacked memories can be constructed by forming memory levels on separate substrates and then stacking the memory levels atop each other. The substrates may be thinned or removed from the memory device levels before stacking, but as the memory device levels are initially formed over separate substrates, the resulting memory arrays are not monolithic three dimensional memory arrays. Further, multiple two dimensional memory arrays or three dimensional memory arrays (monolithic or non-monolithic) may be formed on separate chips and then packaged together to form a stacked-chip memory device.

Associated circuitry is typically required for operation of the memory elements and for communication with the memory elements. As non-limiting examples, memory devices may have circuitry used for controlling and driving memory elements to accomplish functions such as programming and reading. This associated circuitry may be on the same substrate as the memory elements and/or on a separate substrate. For example, a controller for memory read-write operations may be located on a separate controller chip and/or on the same substrate as the memory elements.

It will be recognized that the following is not limited to the two dimensional and three dimensional exemplary structures described but cover all relevant memory structures within the spirit and scope as described herein

Physical Memory Structure

FIG. **2** illustrates schematically a non-volatile memory cell. The memory cell **10** can be implemented by a field-effect transistor having a charge storage unit **20**, such as a floating gate or a charge trapping (dielectric) layer. The memory cell **10** also includes a source **14**, a drain **16**, and a control gate **30**.

There are many commercially successful non-volatile solid-state memory devices being used today. These memory devices may employ different types of memory cells, each type having one or more charge storage element.

Typical non-volatile memory cells include EEPROM and flash EEPROM. Also, examples of memory devices utilizing dielectric storage elements.

In practice, the memory state of a cell is usually read by sensing the conduction current across the source and drain electrodes of the cell when a reference voltage is applied to the control gate. Thus, for each given charge on the floating gate of a cell, a corresponding conduction current with respect to a fixed reference control gate voltage may be detected. Similarly, the range of charge programmable onto the floating gate defines a corresponding threshold voltage window or a corresponding conduction current window.

Alternatively, instead of detecting the conduction current among a partitioned current window, it is possible to set the threshold voltage for a given memory state under test at the control gate and detect if the conduction current is lower or higher than a threshold current (cell-read reference current). In one implementation the detection of the conduction current relative to a threshold current is accomplished by examining the rate the conduction current is discharging through the capacitance of the bit line.

FIG. **3** illustrates the relation between the source-drain current I_D and the control gate voltage V_{CG} for four different charges **Q1**-**Q4** that the floating gate may be selectively storing at any one time. With fixed drain voltage bias, the four solid I_D versus V_{CG} curves represent four of seven possible charge levels that can be programmed on a floating gate of a memory cell, respectively corresponding to four possible memory states. As an example, the threshold volt-

age window of a population of cells may range from 0.5V to 3.5V. Seven possible programmed memory states "0", "1", "2", "3", "4", "5", "6", and an erased state (not shown) may be demarcated by partitioning the threshold window into regions in intervals of 0.5V each. For example, if a reference current, IREF of 2 μA is used as shown, then the cell programmed with Q1 may be considered to be in a memory state "1" since its curve intersects with I_{REF} in the region of the threshold window demarcated by VCG=0.5V and 1.0V. Similarly, Q4 is in a memory state "5".

As can be seen from the description above, the more states a memory cell is made to store, the more finely divided is its threshold window. For example, a memory device may have memory cells having a threshold window that ranges from −1.5V to 5V. This provides a maximum width of 6.5V. If the memory cell is to store 16 states, each state may occupy from 200 mV to 300 mV in the threshold window. This will require higher precision in programming and reading operations in order to be able to achieve the required resolution.

NAND Structure

FIG. **4** illustrates schematically a string of memory cells organized into a NAND string. A NAND string **50** comprises a series of memory transistors M1, M2, Mn (e.g., n=4, 8, 16 or higher) daisy-chained by their sources and drains. A pair of select transistors S1, S2 controls the memory transistor chain's connection to the external world via the NAND string's source terminal **54** and drain terminal **56** respectively. In a memory array, when the source select transistor Si is turned on, the source terminal is coupled to a source line (see FIG. **5**). Similarly, when the drain select transistor S2 is turned on, the drain terminal of the NAND string is coupled to a bit line of the memory array. Each memory transistor **10** in the chain acts as a memory cell. It has a charge storage element **20** to store a given amount of charge so as to represent an intended memory state. A control gate **30** of each memory transistor allows control over read and write operations. As will be seen in FIG. **5**, the control gates **30** of corresponding memory transistors of a row of NAND string are all connected to the same word line. Similarly, a control gate **32** of each of the select transistors S1, S2 provides control access to the NAND string via its source terminal **54** and drain terminal **56** respectively. Likewise, the control gates **32** of corresponding select transistors of a row of NAND string are all connected to the same select line.

When an addressed memory transistor **10** within a NAND string is read or is verified during programming, its control gate **30** is supplied with an appropriate voltage. At the same time, the rest of the non-addressed memory transistors in the NAND string **50** are fully turned on by application of sufficient voltage on their control gates. In this way, a conductive path is effectively created from the source of the individual memory transistor to the source terminal **54** of the NAND string and likewise for the drain of the individual memory transistor to the drain terminal **56** of the cell.

FIG. **4**B illustrates an example of a NAND array **210** of memory cells, constituted from NAND strings **50** such as that shown in FIG. **4**. Along each column of NAND strings, a bit line such as bit line **36** is coupled to the drain terminal **56** of each NAND string. Along each bank of NAND strings, a source line such as source line **34** is coupled to the source terminals **54** of each NAND string. Also the control gates along a row of memory cells in a bank of NAND strings are connected to a word line such as word line **42**. The control gates along a row of select transistors in a bank of NAND strings are connected to a select line such as select line **44**. An entire row of memory cells in a bank of NAND strings can be addressed by appropriate voltages on the word lines and select lines of the bank of NAND strings.

FIG. **6** illustrates a page of memory cells, organized in the NAND configuration, being sensed or programmed in parallel. FIG. **6** essentially shows a bank of NAND strings **50** in the memory array **210** of FIG. **5**, where the detail of each NAND string is shown explicitly as in FIG. **4**. A physical page, such as the page **60**, is a group of memory cells enabled to be sensed or programmed in parallel. This is accomplished by a corresponding page of sense amplifiers **212**. The sensed results are latched in a corresponding set of latches **214**. Each sense amplifier can be coupled to a NAND string via a bit line. The page is enabled by the control gates of the cells of the page connected in common to a word line **42** and each cell accessible by a sense amplifier accessible via a bit line **36**. As an example, when respectively sensing or programming the page of cells **60**, a sensing voltage or a programming voltage is respectively applied to the common word line WL**3** together with appropriate voltages on the bit lines.

Physical Organization of the Memory

One difference between flash memory and other of types of memory is that a cell must be programmed from the erased state. That is the floating gate must first be emptied of charge. Programming then adds a desired amount of charge back to the floating gate. It does not support removing a portion of the charge from the floating gate to go from a more programmed state to a lesser one. This means that updated data cannot overwrite existing data and must be written to a previous unwritten location.

Furthermore erasing is to empty all the charges from the floating gate and generally takes appreciable time. For that reason, it will be cumbersome and very slow to erase cell by cell or even page by page. In practice, the array of memory cells is divided into a large number of blocks of memory cells. As is common for flash EEPROM systems, the block is the unit of erase. That is, each block contains the minimum number of memory cells that are erased together. While aggregating a large number of cells in a block to be erased in parallel will improve erase performance, a large size block also entails dealing with a larger number of update and obsolete data.

Each block is typically divided into a number of physical pages. A logical page is a unit of programming or reading that contains a number of bits equal to the number of cells in a physical page. In a memory that stores one bit per cell, one physical page stores one logical page of data. In memories that store two bits per cell, a physical page stores two logical pages. The number of logical pages stored in a physical page thus reflects the number of bits stored per cell. In one embodiment, the individual pages may be divided into segments and the segments may contain the fewest number of cells that are written at one time as a basic programming operation. One or more logical pages of data are typically stored in one row of memory cells. A page can store one or more sectors. A sector includes user data and overhead data.

All-Bit, Full-Sequence MLC Programming

FIG. **7**A-**7**C illustrate an example of programming a population of 4-state memory cells. FIG. **7**A illustrates the population of memory cells programmable into four distinct distributions of threshold voltages respectively representing memory states "0", "1", "2" and "3". FIG. **7**B illustrates the initial distribution of "erased" threshold voltages for an erased memory. FIG. **6**C illustrates an example of the memory after many of the memory cells have been programmed. Essentially, a cell initially has an "erased" thresh-

old voltage and programming will move it to a higher value into one of the three zones demarcated by verify levels vV_1, vV_2 and vV_3. In this way, each memory cell can be programmed to one of the three programmed states "1", "2" and "3" or remain un-programmed in the "erased" state. As the memory gets more programming, the initial distribution of the "erased" state as shown in FIG. **7**B will become narrower and the erased state is represented by the "0" state.

A **2**-bit code having a lower bit and an upper bit can be used to represent each of the four memory states. For example, the "0", "1", "2" and "3" states are respectively represented by "11", "01", "00" and '10". The 2-bit data may be read from the memory by sensing in "full-sequence" mode where the two bits are sensed together by sensing relative to the read demarcation threshold values rV_1, rV_2 and rV_3 in three sub-passes respectively.

3-D NAND Structures

An alternative arrangement to a conventional two-dimensional (2-D) NAND array is a three-dimensional (3-D) array. In contrast to 2-D NAND arrays, which are formed along a planar surface of a semiconductor wafer, 3-D arrays extend up from the wafer surface and generally include stacks, or columns, of memory cells extending upwards. Various 3-D arrangements are possible. In one arrangement a NAND string is formed vertically with one end (e.g. source) at the wafer surface and the other end (e.g. drain) on top. In another arrangement a NAND string is formed in a U-shape so that both ends of the NAND string are accessible on top, thus facilitating connections between such strings.

FIG. **8** shows a first example of a NAND string **701** that extends in a vertical direction, i.e. extending in the z-direction, perpendicular to the x-y plane of the substrate. Memory cells are formed where a vertical bit line (local bit line) **703** passes through a word line (e.g. WL**0**, WL**1**, etc.). A charge trapping layer between the local bit line and the word line stores charge, which affects the threshold voltage of the transistor formed by the word line (gate) coupled to the vertical bit line (channel) that it encircles. Such memory cells may be formed by forming stacks of word lines and then etching memory holes where memory cells are to be formed. Memory holes are then lined with a charge trapping layer and filled with a suitable local bit line/channel material (with suitable dielectric layers for isolation).

As with planar NAND strings, select gates **705**, **707**, are located at either end of the string to allow the NAND string to be selectively connected to, or isolated from, external elements **709**, **711**. Such external elements are generally conductive lines such as common source lines or bit lines that serve large numbers of NAND strings. Vertical NAND strings may be operated in a similar manner to planar NAND strings and both SLC and MLC operation is possible. While FIG. **8** shows an example of a NAND string that has 32 cells (0-31) connected in series, the number of cells in a NAND string may be any suitable number. Not all cells are shown for clarity. It will be understood that additional cells are formed where word lines 3-29 (not shown) intersect the local vertical bit line.

A 3D NAND array can, loosely speaking, be formed tilting up the respective structures **50** and **210** of FIGS. **5** and **6** to be perpendicular to the x-y plane. In this example, each y-z plane corresponds to the page structure of FIG. **6**, with m such plane at differing x locations. The (global) bit lines, BL**1**-*m*, each run across the top to an associated sense amp SA**1**-*m*. The word lines, WL**1**-*n*, and source and select lines SSL**1**-*n* and DSL**1**-*n*, then run in x direction, with the NAND string connected at bottom to a common source line CSL.

FIGS. **9-12** look at a particular monolithic three dimensional (3D) memory array of the NAND type (more specifically of the "BiCS" type), where one or more memory device levels are formed above a single substrate, in more detail. FIG. **9** is an oblique projection of part of such a structure, showing a portion corresponding to two of the page structures in FIG. **5**, where, depending on the embodiment, each of these could correspond to a separate block or be different "fingers" of the same block. Here, instead to the NAND strings lying in a common y-z plane, they are squashed together in the y direction, so that the NAND strings are somewhat staggered in the x direction. On the top, the NAND strings are connected along global bit lines (BL) spanning multiple such sub-divisions of the array that run in the x direction. Here, global common source lines (SL) also run across multiple such structures in the x direction and are connect to the sources at the bottoms of the NAND string, which are connected by a local interconnect (LI) that serves as the local common source line of the individual finger. Depending on the embodiment, the global source lines can span the whole, or just a portion, of the array structure. Rather than use the local interconnect (LI), variations can include the NAND string being formed in a U type structure, where part of the string itself runs back up.

To the right of FIG. **9** is a representation of the elements of one of the vertical NAND strings from the structure to the left. Multiple memory cells are connected through a drain select gate SGD to the associated bit line BL at the top and connected through the associated source select gate SDS to the associated local source line LI to a global source line SL. It is often useful to have a select gate with a greater length than that of memory cells, where this can alternately be achieved by having several select gates in series (as described in U.S. patent application Ser. No. 13/925,662, filed on Jun. 24, 2013), making for more uniform processing of layers. Additionally, the select gates are programmable to have their threshold levels adjusted. This exemplary embodiment also includes several dummy cells at the ends that are not used to store user data, as their proximity to the select gates makes them more prone to disturbs.

FIG. **10** shows a top view of the structure for two blocks in the exemplary embodiment. Two blocks (BLK**0** above, BLK**1** below) are shown, each having four fingers that run left to right. The word lines and select gate lines of each level also run left to right, with the word lines of the different fingers of the same block being commonly connected at a "terrace" and then on to receive their various voltage level through the word line select gates at WLTr. The word lines of a given layer in a block can also be commonly connected on the far side from the terrace. The selected gate lines can be individual for each level, rather common, allowing the fingers to be individually selected. The bit lines are shown running up and down the page and connect on to the sense amp circuits, where, depending on the embodiment, each sense amp can correspond to a single bit line or be multiplexed to several bit lines.

FIG. **11** shows a side view of one block, again with four fingers. in this exemplary embodiment, the select gates SGD and SGS at either end of the NAND strings are formed of four layers, with the word lines WL in-between, all formed over a CPWELL. A given finger is selected by setting its select gates to a level VSG and the word lines are biased according to the operation, such as a read voltage (VCGRV) for the selected word lines and the read-pass voltage (VREAD) for the non-selected word lines. The non-selected fingers can then be cut off by setting their select gates accordingly.

FIG. **12** illustrates some detail of an individual cell. A dielectric core runs in the vertical direction and is surrounded by a channel silicon layer, that is in turn surrounded a tunnel dielectric (TNL) and then the charge trapping dielectric layer (CTL). The gate of the cell is here formed of tungsten with which is surrounded by a metal barrier and is separated from the charge trapping layer by blocking (BLK) oxide and a high K layer.

Array Structure Defects

Memory arrays such as those described above are often subject to various defects, such as broken or leaky word lines and bit lines. A number of techniques for the determining of, and the dealing with, these sorts of problems are presented in the following US patent publications: 2012-0008405; 2012-0008384; 2012-0008410; 2012-0281479; 2013-0031429; 2013-0031429; and 2013-0229868. The following present a number of additional techniques in the context of the sort of 3D memory structures described above with respect to FIGS. **9-12**. Although these techniques are particularly applicable to such structures, many of them are more generally applicable, such as to 2D NAND and other array structures.

Word Line to Word Line Shorts, Same Block: AC Stress Mode

This section considers shorts between word lines between of the same block, whether in a 2D array or a 3D. In a 3D arrangement, such as illustrated in the FIGS. **9-12**, a number of word lines are stacked on top of each other, with an oxide layer in between the pairs of word lines to act as an insulating layer. If this oxide layer is not deposited uniformly or, due to some contamination, is thinner than the target values, the oxide can fail at some point, leading to a short. The stress mode described in this section can accelerate the failure to occur as part of a defect determination process. FIG. **13** can help to illustrate the problem.

FIG. **13** is a side view of a block, similar to FIG. **11**, but with some of the features relevant to the current discussion highlighted. As before, four fingers of a block are shown with some of the global bit lines, source lines and so on running across the top. More specifically, the two arrows to the right show two of the gaps between word lines that would be filled with oxide. (As discussed further below, these oxide layers are also between select gate lines, as shown in the second finger.) The potential word line to word line shorts would then be across these oxide layers, such as illustrated at the right most finger.

A number of references cited above present techniques for determining such shorts. Typically, these use a DC stress applied to the word lines. This section uses an AC stress that, in the exemplary embodiment, is applied to the odd and even word lines; for example, while toggling the even word lines to a high voltage level VH, the odd word lines are toggled to a low voltage level VL. (More generally, this can be done to with two sets of word lines, such as a portion of the word lines, where the two sets have at least one word line from each that is adjacent.) VH is taken as a high voltage level such that the Delta (VH-VL) is high enough to stress the oxide in between adjacent WLs. For example, VH can be as high as 20V to reflect the sort of word line stress levels seen due to program and erase voltage levels used on the device. VL can be the low level (ground or VSS) on the device. This toggling is illustrated schematically in FIG. **14**, that shows the voltage levels being applied to the two sets of word lines.

As shown in FIG. **14**, the even word lines alternately have the VL and VH levels applied, with the odd word lines being similarly toggled but with the phase reversed. These waveforms can be based on a number of settable parameters, including the time period of the VII level, defined by parameter Th, and the time period of VL level, defined by parameter TL. Th and TL can be adjusted to have Delta (VH-VL) between adjacent WLs for a fixed duration. In case the rise or the fall time of VH and VL differ, the parameters Th and TL can be adjusted to achieve a desired duration of the stress in each cycle. The number of loops of this AC stress can also be parameterized (loop count parameter) to prevent any kind of over kill or under kill. Although the exemplary embodiment applies these levels along the sets of even and odd word lines, as noted above this can be done on other subsets that have one or more adjacent word lines within the block or finger.

FIGS. **15** and **16** are an exemplary flow for the on-chip stress (FIG. **15**) and defect determination (FIG. **16**) process. The process begins with the test mode entry at **1501**. This will typically be part of an initial test process, done before a device is packaged or shipped, to determine defective die, but also can also be done after the device has been in operation for some time, where this could them be implemented at the system level and with possible adjustments to the parameters (see **1509**). At **1503** it is determined whether to just do a single block test or multiple, even all, blocks. In other embodiments, partial block test can also be done. The block (**1507**) or blocks (**1505**) are then selected. The stress is then applied at **1509**, which show the parameters involved. The idea is to apply the stress to break any weak spots during the test, basically forcing the issue, rather than wait for these failures to occur during operation.

After the stress operation comes the defect determination operation. A number of techniques for this are described in US patent publications: 2012-0008405; 2012-0008384; 2012-0008410; 2012-0281479; 2013-0031429; 2013-0031429; and US2013-0229868. For the exemplary flow here, the stressed block or blocks are programmed and then can be read back to check for failures. The flow picks up at **1601** with a program operation, the status of which is monitored at **1603**. If the write operation fails (**1605**), the bad block or blocks are marked as such and not used. Alternately, this (as well as **1611** below) could also be done at the word line level. If the program operation passes, the detection can then move on to a read operation at **1607**. The read status is monitored at **1609** by, for example, comparing the data as read back with the data that was originally programmed. If the read status comes back as a fail, then the bad block or blocks can be marked (**1611**) as such. If the tested blocks pass (**1613**), the test mode can then exited (**1615**).

For the defect determination operation, determining whether a program operation has failed can be based on a word line failing by exceeding maximum number of program loops or on the read operation coming back with a failed bit count exceeding a limit. The data written and read back for this process can be a random pattern, either predetermined or not. In either case, the comparison of the read back data can be based on a comparison of the data pattern that was to be written, each by maintaining a copy or because it is based on a known, predetermined pattern. Rather than a direct comparison, it can alternately (or additional) be based on ECC and whether the data can be extracted within the ECC capabilities.

FIG. **17** is a schematic representation of some of the elements on the memory chip that are involved in this process. A number of different embodiments are possible, but shows some of the basic elements. The array **1701** and its associated decoding and sensing circuitry can be of the BiCS or other 3D variety, but is here shown in more of a 2D

sort of representation for simplicity. The memory circuit has a set of JO pads **1703** for commands, addresses and data transfer, which can then be passed on to command and block/page address registers (**1705**, **1707**). An oscillator **1709** can be used with the clock generator to provide needed clock signals. A finite state machine (FSM) and Sequencer block **1713** represented to on-chip control logic that controls the various drivers **1715** for the array **1701**. References voltage and current generators **1721** supply the various reference levels, including those supplied to the charge pump circuits block **1723**, which generates the high level VH applied to the word lines in the intra-block AC stress mode of this section. A pulse generation circuit, here schematically separated out as block **1725**, will then supply the stress level to the selected block when enabled by mode select signal, as schematically represented by the control signals to the control gates of the high voltage switches **1727** and **1729**.

For any of the variations, these techniques can be effect for detecting current or incipient word line to word line shorts within a given block, without over identification, improving the identified defective parts per million (DPPM) value. This process can be implemented as a built in self-test (BIST) process that can help in reducing test times and also gives the option to perform the stress at the system level at other times, such as before performing a block.

Word Line to Word Line Shorts, Adjacent Blocks

In most non-volatile array structures, the issue of word line to word line shorts is traditional an issue for word lines within a common blocks. In some structures, such as the sort of 3D BiCS-type structure described above with respect to FIGS. **9-12**, word lines from different, but physically adjacent blocks can be in close proximity. This section looks at the detection of such inter-block word line to word line shorts. As can be seen from FIG. **9**, where the two shown fingers are from different blocks, word lines on the same vertical level from different blocks can be closely placed, separated by a distance that will decrease as device scales continue to shrink.

Word line to word line shorts from adjacent blocks can manifest themselves as erase disturbs, program disturbs, or both. For example, consider the case where a block X is already in a programmed state and unselected for erase, but an adjacent block, block X+1, is selected for erase. In case of a short between a word line WLn of block X and word line of block X+1, when applying the erase voltage level to block X+1, the erased bias level will transfer across the shorted word line WLn to block X, causing some degree of erase there so that block X will lose already programmed data, resulting in erase disturb there. Similarly, if block X is already programmed and meant to be unselected for additional programming while the adjacent block X+1 is selected for programming, when a programming pulse is applied along WLn in block X+1, this will be transferred across the short. WLn in block X be get over programmed and WLn of block X+1 will see high loading, resulting in program disturb there.

FIG. **18** is a top level, top down diagram of how blocks are paired and placed next to each other in the exemplary embodiment. The four (in this embodiment) fingers of block **0 1801** are at top, with each word line in the stack connected to a corresponding terrace level that in addition to being to the left of the fingers of block **0**, are also in proximity to the fingers of block **1 1803**. To the left and right are the respective transfer gates for block **0** (**1805**) and block **1** (**1807**). The bit lines then run up and down in this view, connecting the NAND strings running into the page to the associated sense amps. The arrows show some of the examples of where inter-block word line shorts can occur, such as between a word line of one block and the other blocks terrace region **1811** or between the word lines of adjacent fingers of different blocks **1813**. Also this discussion is given in the context of the BiCS type structure, it can also be applied to other structures where word lines of one block are in proximity to the word lines of another block.

The process for determining inter-block word line to word line shorts between adjacent blocks can again use a stress operation and a detection operation, where the stress phase can be an AC or DC stress. The high level VH is again a high voltage level such that the Delta (VH-VL) is high enough to stress the oxide or weak defect in between adjacent word lines. VH can be as high as 20V to correspond to the stresses the device will see during erase and write operations. The low level VL is similarly such that the Delta (VH-VL) is high enough to stress the oxide or weak defect in between adjacent word lines, where VL can be as low as the low on-chip level of VSS or ground. After the stress mode, whether AC or DC, a word line to word line leakage determination can then be done to check pass/fail status using decoder circuit, modified as appropriate.

The inter-block stress between word lines can be applied in a number of ways. For example, a differing bias can be applied on the word lines from even and odd blocks. In one embodiment, this could be applying a high voltage to all word lines on even blocks and a low voltage to odd block, or vice versa. This will create a stress between word lines of physically adjacent blocks to be able to weed out defects at time at test time. This stress can be applied to two or more physically adjacent blocks at a time.

FIG. **19** schematically represents the word line levels as applied to even and odd block under this arrangements, where the word lines on even blocks are set high and the odd word lines are set low, where this can also be done the other way around. These levels can then be held for some duration in a DC stress mode, or toggled over a number of cycles. These levels can then be applied to by the corresponding decoding circuitry and transfer gates to the word lines according to this pattern.

In an alternate stress mode, a stripe pattern can be applied, either as DC or toggled for a AC mode, to the word lines of adjacent blocks, where one pattern is applied for one block and the pattern reversed for the adjacent block. For example, on the word lines of even blocks, voltage levels can be applied in an alternating high-low pattern, as shown in FIG. **20**, with the inverted low-high pattern on the odd blocks. This stress can again be done a pair of blocks at a time or multiple selected of blocks.

A post-stress detection sequence can then follow. The exemplary embodiment for a detection sequence uses an erase disturb test to detect the word line to word line shorts from adjacent blocks. This can be done by erasing all blocks of an array. For all blocks that pass the erase, a program follows to see whether blocks program correctly. For example, random data can be programmed on all blocks of the array, read back, and compared against the expected data to check for any program disturb. A block N can then be erased, after which an adjacent block N+1 is checked for any erase disturb.

FIG. **21**A is a schematic representation of some of the elements on the memory chip, similar to FIG. **17** and with the corresponding elements similarly numbered (i.e., array **1701** is now array **2101** and so on). To enable the inter-block word line to word line stress mode, a corresponding stress mode select signal is sent from the on-chip control logic to

the charge pump, which in turn can supply the needed high voltage level to the control gate drivers in block **2115**. These drivers can then supply the high and low stress voltages CGUE and CGUO to the even and odd blocks. FIG. **21**B gives more detail on the decoding circuitry and its connections to the array for the exemplary embodiment, where the even block terraces are to the left and the odd block terraces to the right.

For any of the variations, the test mode of this section can consequently be used to catch word line to word line shorts between word lines of adjacent blocks. By using an AC version of the sort of alternating high and low voltage pattern illustrated with respect to FIG. **20**, this can concurrently apply the sort stress described in the preceding section for word line to word line shorts within a block, helping to reduce test time. For the voltage patterns of either of FIG. **19** or **20**, the separate biasing of neighboring block is used, which is not conventional in NAND memory decoder designs, but which can be further enhanced for parallelism by using opposite alternating patterns on facing blocks. As with the other defect determination methods described here, the detection of adjacent blocks' erase disturb, program disturb, or both can help to avoid the corrupting of user data.

Select Gate Shorts

The preceding two sections looking at defects that could to shorts between word lines. Going back to the structure shown FIGS. **9** and **10**, this has multiple select gates on both the source and drain sides, so that in each finger there will be oxide layers between the end most word lines and the adjacent select gates lines (on both the source and drain ends) and also between the multiple select gate lines themselves. There will also be oxide layers between the corresponding select gate layers in different, adjacent blocks and different adjacent finger. Also, as the select gates in the exemplary embodiment have tunable thresholds—that is, they are programmable—they will also be subjected to high voltage levels. Consequently, select gate to select gate and word line to select gate shorts can occur.

Although similar to the word line case, the select gate structures have some differences that can be illustrated with respect to FIG. **22**. FIG. **22** is an oblique view of a simplified version of one block with four fingers, where each NAND string only has two memory cells and only a single select gate at either end. Considering the rightmost finger, the pair of word lines **2221** and **2225** are between the select gate lines **2201** and **2211**. As discussed above, the word lines of a given level from different fingers of the same block are connected together, as shown at **2223** for word line **2221**, along the terrace region. The select lines of the fingers, however, are separate, so that the fingers can individually be selected. Consequently, the select gate lines at **2201**, **2203**, **2205**, and **2207** are independent controllable, lacking the sort of connection between fingers that **2223** effects for the word lines. The simplified drawing of FIG. **22** shows only a single select gate on either end, but when there are multiple select gates, these are typically operated in unison as they perform a common select function.

In this structure, the consequences of a select gate to select gate short can be more severe that a word line to word line short, as can be illustrated by comparing FIGS. **23** and **24**. Both of these show four NAND strings from four different fingers of the same block. A given word line connects the gates across the fingers, while the select gates are individually controllable (where, in these figures, only a single select gate is shown on each of the source and drain sides). Referring to FIG. **23**, a short between two word lines (such as WL**0** and WL**1**) would cause all of the pages of data on these word lines of the block to be lost. Referring now to FIG. **24**, if instead there is a short between select gates of different finger (such as SGD**0** and SGD **1**), the corresponding NAND strings cannot be independently accessed, causing the loss of a much larger number of pages. As with the word line case, there are several types of shorts to consider. There can be select gate to select gate shorts in the same finger of the same block; across different fingers of the same block; and between blocks. There can also be shorts between the end most word lines and adjacent select lines. (A further possible short to the interconnect or local source line, affecting both word lines and select gate line, is discussed in the next section.) For each of these mechanism, a corresponding stress can be applied followed by a detection operation. Either an AC or DC stress can be used and can be done independently or, when appropriate, can be incorporated into the word line test operations. In some cases, such as the same block, different finger case, that is not available for the word line to word line case as illustrated in the FIG. **24**, this would need to be an independent stress operation. Further, in addition to the stress and detection operations described herein, those presented in US patent publications: 2012-0008405; 2012-0008384; 2012-0008410; 2012-0281479; 2013-0031429; 2013-0031429; and US2013-0229868 can also be adapted to the select gate, and select gate-word line, cases.

The biasing voltages for select gate-select gate stress can be different compares to word line case, both to reflect the different levels that are used on these different transistors and also on the different decoding options that may be available on the selects of the device. Although the drain and source select transistor sets are typically operated respectively as a single drain and source transistor, if some or all of the elements of each set of (in this example) four can be individually biased, if the needed select gate decoding is available, this can be used for applying a strip pattern, either AC or DC, between the select gates lines of the same finger, different fingers of the same block, adjacent fingers of different blocks, or some combination of these. In the case where the set of select gates are share a gate contact (or, similarly, for only a single such gate), these will all have the same bias level and the stress is only applied between fingers (whether the same or different blocks). In an intermediate sort of situation where, say, one source select gate line ("SGSB") can independently biased while the other three (SGS) are connected together to have the same control, several stress options are available, such as: SGSB-SGSB (adjacent fingers); SGS-SGSB shorts (same finger, as for a word line to word line short); and SGS (either of 3 SGS of one finger)-SGS (either of 3 SGS of adjacent finger).

Word Line to Local Source Line Shorts

In an array with a NAND type of architecture, the NAND string of a block are typically connected to a common source line, as shown at **34** of FIG. **5**. The source lines of multiple blocks, even all of the blocks of a die, often share such a common source line. Referring to the 3D structure of FIG. **11**, in the arrangement shown there, between each of the fingers of the same block a local common source line CELSRC (or local interconnect, LI) runs up to connect the source side of the NAND strings to one or more global common source lines (not shown in FIG. **11**) running across the top the structure. As can be seen in FIG. **11**, this places the word lines (and select gate lines) in proximity to this local source line interconnect, leading to the possibility of shorts across the intervening oxide,

During an erase operation in this sort of structure, the LI (CELSRC) will couple to the high erase voltage, while the

word lines (and any dummy word lines) are low (0V) and the select gate lines can be either driven or floated to prevent them from being erased. In case of an LI to word line short, the erase voltage will be droop and the device may not be able to successfully erase the word lines. This defect can also cause read and program operations to fail. This is block level failure. This section looks at methods for determining such defects at test time.

In 2D NAND devices, there are often modes that apply a stress (high voltage) on word lines, while keeping the source line low (close to 0V), but this sort of stress mode can degrade the characteristics of the memory cells, leading reliability and endurance concerns. Additional, in a typically 2D arrangement, the metal line of a source line does not run next to word lines, so that the failure mode considered in this section is more specific structures, such as the BiCS array, that have this sort of lay out.

In the exemplary embodiment, a high voltage (~VH) is applied on the local common source line of the block (LI or CELSRC), and lower level is applied to the word lines, including any dummy word lines, with the select gate lines either driven or floated. Here VH is a voltage level such that it is high enough to break weak oxide between LI and any word lines, but small enough such that reverse bias diode between the CPWELL (p+) and the CLESRC (n+) region does not break down; for example, in the exemplary embodiment this can be on the order an erase voltage, say 20V. The low level can be taken from among the low levels on the chip, such as VWL, VSG, or VL. Both VWL and VSG are also voltage levels close to VSS or VSS, with values such that they will not stress the memory cells, so that endurance and reliability are not adversely affected. In this example, VSG is VSG voltage is mainly a biasing voltage to turn on the NAND string during read/program/erase operations and VWL can be mainly be VCGRV (control gate read-verify) level, where VCGRV voltage can go as low as close to 1V in an exemplary embodiment. The CPWELL level can be set at range of values between the high and low levels, as long as combinations with the other voltages does not break down the reverse bias diode between the CPWELL (p+) and the CLESRC (n+) region.

This arrangement of bias levels will stress weak oxide depositions, whether due to contamination or other defect, in the region between the word lines and the local source line interconnect in order to bring about a short. This can cause the defect to manifest itself at test time, rather than once the device is in use. As high voltage levels are not placed on the word lines, the cells will not be stressed, avoiding adverse effects for reliability and endurance. FIG. **25** again shows a side view of the exemplary embodiment with an overview of the applied voltages.

In FIG. **25** the word line to local source line stress is applied in the circled region. The low voltage levels on the leftmost finger and the high voltage on the vertical source line interconnect. The arrows illustrate particular examples of locations of stress between a word line and the source interconnect line where leakage could occur.

FIGS. **26** and **27** are an exemplary flow of a test mode to screen local interconnect (LI) to word line physical shorts, similar to FIGS. **15** and **16** above for the intra-block word line to word line case. The test mode can either be done as part of an initial test process or as a system level countermeasure where, after some number of cycles and before erasing any blocks, the test can be done to catch any bad blocks before programming in any data. FIG. **26** is the stress phase and starts at **2601** with entry into the text mode. At **2603** the determination is made as to whether a single block or multiple blocks are to be tested and at **2605** or **2607** the corresponding select signals are sent out. The stress voltages are then applied at **2609**. As with the word line to word line cases, the duration of the stress is a settable parameter, as can be the various levels.

FIG. **27** illustrates an exemplary embodiment for the detection phase, where many of the specifics are as discussed with respect to FIG. **16**. This begins with a program operation at **2701**, which is monitored at **2703**. If the program fails for any blocks, these are marked bad (**2705**). If the program operation passes, a read can then be performed at **2707**. The read status is then monitored at **2709** and the result can be checked by, for example, comparing the data as read back with the data as stored. Any blocks that fail are marked bad at **2711**. If the read operation passes, an erase can then be performed at **2713**. This is monitored at **2715**, with any failing blocked marked at **2717**. If the erase is passed (**2719**), the test mode is then exited (**2721**).

FIG. **28** is a schematic representation of some of the elements on the memory chip, similar to FIGS. **17** and **21**, and with the corresponding elements similarly numbered. To enable the inter-block word line to word line stress mode, a corresponding stress mode select signal is sent from the on-chip control logic to the charge pump, which in turn can supply the needed voltage levels for the word lines and select gate lines to the drivers in block **2115**. The usual levels for CELSRC are supplied through switch **2825**, with high voltage VH from the charge pump can be supplied through a high voltage switch **2827**, that also can get its control gate voltage from the charge pump.

Bit Line to Low Voltage Signal Shorts

In the memory array, global bit lines span the structure connecting the memory cells to the sense amplifies used in sensing operations. This shown above in FIG. **18**, for instance, where each NAND string is connected to a bit line, and each bit line is connected to a NAND string in finger. The sense amps are then located on the periphery of the array. In the exemplary embodiment, and for simplifying this discussion, each sense amp connects to a single bit line. In other embodiments, where less than all of the bit lines are sensed at once (say, every other or every fourth bit line being sensed concurrently), multiple bit lines (such as 2 or 4) are associated with each sense amp. As the spacing of bit lines is typically smaller than the area needed by the sense amp circuits, the sense amp circuits are often staggered relative to one another in the chip's layout. This means that one bit line may be adjacent to the sense amp associated with another bit line.

During erase operations in some memory circuit designs, such as the exemplary BiCS type embodiment illustrated with respect to FIGS. **9-12**, during an erase operation bit lines will couple to the erase voltage (which can be in the ~20-24V range), taking the bit lines to a high voltage. The sense amplifiers generally operate at lower voltages, such as VSS (0V), the high logic level (VDD, in the ~2-3V ranges), and, in some embodiments, a somewhat higher sense amp level used in pre-charging bit lines for sensing operations (VDDSA, in the ~4-5V range). Due to the proximity of the bit lines at high voltages with sense amp circuitry at relatively low voltages, the bit lines can short to the adjacent lower voltage circuitry. In case of such a bit line to low voltage signal short during device operations, the erase voltage (VERA) may droop and circuit will not be able to successfully erase selected blocks. Even if the bit line is repaired, this fault can still cause erase failure. Further, as

the bit lines are global, this will be global defect for the portion of the array (typically the entire plane) spanned by the defective bit line.

FIG. **29** is a schematic representation of the situation in the BiCS context. A side view of several fingers of the array are shown, a pair of adjacent bit lines BLn **2905** and BLn+1 be explicitly shown at top. In an erase operation, the CPWELL is biased to the high erase voltage Vera. The unselected NAND strings will transmit the high voltage to these bit lines. Each of bit lines BLn **2905** and BLn+1 **2915** are respectively connected an associated sense amp SAn **2901** and SAn+1 **2911** though a switch **2903** and **2913**, where sense amps are connect to the switches by an "internal" part of the bit line BLIn, BLIn+1. During erase, the switches **2903**, **2913** are shut off, protecting the sense amps' circuitry and keeping the high voltage contained. Due to the layout, however, one bit line (such as BLn) may be adjacent to the BLI portion, or other sense amp elements, associated with another bit line (such as BLn+1), resulting in a short as represented at **2909**. Consequently, for any erase operation the high voltage Vera on the CPWELL will drain off through bit line BLn **2905**. If the charge pump supplying the erase voltage in not able to keep up with this leakage, Vera will not be maintained and the erase operation will fail. Although illustrated for a 3D arrangement, this problem can also show up in other architectures, including in 2D arrays. As in the preceding sections, to detect such defects, a test can mimic the stress involved to force incipient failures and then perform a detection phase to see whether the circuit is able to perform the needed operations.

In the stress operation, one or more bit lines are set to the high voltages, while one or more adjacent internal bit lines (BLI) and/or associated sense amp circuits are set to a lower sense amp voltage. Although this can be done by applying the high level directly to the bit lines, most memory circuits typically do not include such connections. Consequently, the exemplary embodiment establishes the high bit line voltage as these would come about through normal circuit operations, namely from the CPWELL. This can be done in a 1-plane erase failure stress mode by applying the high voltage to the P-well (e.g., bias ~20-24V) and having all blocks unselected, so the bit lines will couple through the NAND strings to the high voltage, while concurrently setting the sense amp nodes to the (relatively) low sense amp voltages (VDDSA/VDD/VSS). The stress can be applied in either a DC mode for some duration, or in an AC by applying some number of pulses of given durations to the P-well.

Any resultant bit line short would drain off the high voltage from the P-well, reducing its ability to effect an erase as the charge pump supplying the high voltage may not be able to keep up. This is used in the detection operation of the exemplary flow by reducing the drive capability of the charge pump and determining whether the erase voltage can be held. For example, the pump clock can be set at the slower end of its range and the Vera voltage being measure internally to detect any droop due to a bit line high voltage-sense amp node low voltage defect/leak. A comparator circuit can use a reference voltage for compassion with the Vera level. Alternately, the pump clock can be set at its slowest and an auto-detect built in self-test (BIST) mode can be based on the charge pump's ON time being, due to the leakage, longer than typical

This technique will stress the bit line to low voltage node, accelerating the failure of defects, but without overly stress the memory cells of the array. FIG. **30** is similar to FIG. **29**, but marked with some of the voltage levels involved in the stress phase. As discussed above, VH is a voltage level high enough to break weak oxide between the bit lines and the low voltage nodes, but small enough such that there no degradation of the select gates or any other circuit elements. LV is the relatively low voltages used by the sense amp nodes (e.g., VDDSA/VDD/VSS).

In the exemplary embodiment, the blocks are all unselected, with the NAND strings conducting between the P-well and bit lines. This stress mode is chosen so as to minimize effects on cell characteristics to avoid negative effects on reliability and endurance. The stress mode can also be used at system level by applying the bit line to low voltage node stress to catch full plane erase issues prior to programming data into the array. The stress can be applied at time **0** or after some numbers of program-erase cycles.

FIG. **31** is an exemplary flow for these test processes, beginning by entering the test mode at **3101**. The stress is applied at **3103**, where the duration and voltage levels can be settable parameters. If an AC stress mode is used, this will applied for N loops **3105**, where N can also be a settable parameter. In the exemplary detection flow, the pump clock is set to its slowest (**3107**), all of the blocks in the plane are unselected and the Vera is supplied to the P-well from the pump (**3109**), and value of Vera on the P-well is measured (**3111**). At **3113** the P-well level compared against a reference level and if it has drooped, a flag can be set according (here, as high at **3119**) and this status can be reported out at **3121**. If Vera is maintained, the flag is set (in this example) as low at **3115** and the status reported out at **3117**.

FIG. **32** is a schematic representation of some of the elements on the memory chip, similar to (and similarly number to) FIGS. **17**, **21** and **28**, for implemented the test mode of FIG. **31**. During the detection phase as enabled by the stress mode select signal, while the detuned charge pump drives the P-well the switch SW **3221** supplies the P-well level to the comparator **3223**. This can be stepped down by the step down circuit **3225** for comparison to a reference level Vref. The output of the comparator **3223** can then be converted to a digital value at AID **3227** and stored in the register **3229** as the Flag value, where the register **3229** can receive a clock signal from the FSM & Sequencer block **3213** to latch the flag value. The flag value will indicate if the plane failed, and put out at the I/O page **3231** (separated out here form the other pads for illustrative purposes).

AC Stress Method for Bit Line-Defects

The preceding section considered shorts between bit lines and low voltage circuitry related to the sense amps. This section considers techniques for determining other bit line defects, including bit line to bit line shorts, bit line to "memory hole" defects, and resistive bit lines. An AC stress mode is used to accelerate the bit line defects, where this can be applied as a global stress concurrently across all blocks.

The memory hole defect can be illustrated with respect to FIGS. **9** and **12** for a 3D structure. As shown in FIG. **9** the bit lines (BL) run across the top of the structure in the x direction, with the NAND strings formed into the memory in holes that run down into the structure in the z direction. The global bit line then connects to n+ region (see FIG. **12**) of a corresponding NAND string at the top; or, rather, it is supposed to connect to the bit line unless there is a memory hole open issue with the various resistive contact involved. An analogous defect can also occur in 2D structures, such as at the location indicated by reference number **56** in FIG. **5**, but the additional complexity of the BiCS type structure tends to make them more prone to this sort of problem. This section looks at a stress and detection to catch the memory mole open defect, as well other other bit line related defects.

FIG. **33** shows a view of an array structure similar to FIG. **30**, but with an example of the contact in question circled. Specifically, bit line BLn is connected to the left-most NAND string of each of the three shown fingers. These bit line contacts to the n+ region for memory hole can be very resistive, causing the string to be very resistive and unable to accurately sense the bits along the string current. Such open/resistive contacts can happen on multiple block locations in the array. A bit line to bit line DC stress is typically not sufficient to accelerate this problem so that it can be caught at time **0** (test time); and a normal read will also not readily detect such resistive bit lines. The following instead uses an AC stress to check for bit line to memory open contacts and other bit line defects.

More specifically, an AC voltage is applied to bit lines, toggling between a high voltage (VH) and a low voltage (VL) (or the other way around), to accelerate the defect. The stress can be applied to single bit lines, groups of multiple bit lines, or as a global stress applied to all of the memory hole contacts concurrently. In the exemplary embodiments, these levels are applied to the bit lines from the sense amps to drive the bit line side of the contacts. The other side of the contact can be driven by the setting the CPwell to a level VX through the memory hole, where VX can be 0V or a negative voltage (e.g. ~−2V), where this level can be set on the well that is then left to float while the AC voltage toggles on the bit line side. Additionally, if the AC stress is applied out of phase to physically adjacent bit lines, this can be concurrently be used as a stress for detecting bit line to bit line shorts. This use of an out of phase AC stress is illustrated in FIG. **33** where SAn is driving BLn at VH when SAn+1 is driving BLn+1 at VL, and vice versa. For example, this can be used at a global bit line level by concurrently driving the even and odd bit lines out of phase with one another.

An exemplary set of waveforms for the AC voltage applied to the bi lines is represented in FIG. **34**. The cycle time Tx and the number of cycles N can both be settable parameters that can be determined as part of device characterization. The VH value applied by the sense amps can be a regular operating value of sense amp, such as VDDSA (~3-4V), or a higher level used for this purpose. VL can be 0V or VSS. The level on the CPwell can be set at VX=0V or even a negative level such as −2V and then left to float while the stress is applied.

After the stress phase, a detection phase follows. The detection check for bit line to memory hole open contacts can be done across all blocks, checking one block at a time. All of the blocks to be checked are erased and then read against the expected erased block result (an all FF data paten). An open contact would correspond to a string not returning the expected result. For the detection, the read can be modified to pre-charge the bit lines to a higher voltage and the sensing time reduced to more effectively detect any resistive strings. This can also be followed by a check for bit line to bit line shorts by programming and reading back data; for example, a random data pattern can be written in and read back, similar to the process described for word line to word line shorts as described with respect to FIG. **16**. Any defective structures can be marked as bad, where this can be done at the string or block for a bad contact or at the column level or bit line level for bad bit lines.

The techniques of this section have system level advantages. At the system level, highly resistive memory holes can be detected as described above and marked as "unrepairable" or replaced with redundancy or extra strings (local columns) available at the block level. The AC stress mode and detection be performed as part of a built in self-test process as well as being applied at system level to accelerate bit line to bit line short, resistive bit lines and marking of bad columns.

FIG. **35** is a schematic representation of some of the elements on the memory chip, similar to (and similarly number to) FIGS. **17**, **21**, **28**, and **32**, for implementing the AC bit line test mode of this section. During the detection phase as enabled by the stress mode select signal, the block **3531** is the AC voltage generator for the bit lines. This will apply the VH levels through the switches **3533** and **3535** respectively to the even and odd bit lines with the time as controlled by the period Tx. For detection, the sense amps in the peripheral circuitry can pre-charge the bit lines to higher voltages and reduce sensing time to detect any resistive memory hole.

Improved Techniques for Detecting Broken Word Lines

This section returns to the consideration of word line related defects, and more specifically to techniques for the detection of broken word lines. As with the earlier sections, the techniques present here can be combined with the various features found in US patent publication number 2012-0008405 and other references cited above.

Considering the problem of broken word lines in the context of a BiCS memory as illustrated in FIGS. **9-12**, a word line can be broken in several ways. A first type of broken word line relates to the routing between the word line decoding circuits and the portion of the word lines along the control gates. Referring to FIG. **10**, the terrace region to the sides of the memory columns are etched to form a stair case area for the lines from the decoding circuitry to hook up. Along the stair case area of the terrace region, the lines from the decoding circuitry run over the structure and an interconnect drops down to connect to the corresponding word line. This is illustrated in FIG. **36** where a number of word lines (WLs) are separated by oxide layers, where the memory cell region is to the right of the broken lines and the terrace region to the left. To connect to the corresponding step, the oxide (and other layers above the word line) need to be etched away. In the case of an under-etch, the interconnect connection to the word line will be highly resistant. An analogous defect can occur in 2D memory structures, but the particulars of the BiCS and similarly structures are particularly prone to this type of defect.

A second type of word line break is illustrated in FIG. **37** where, due to non-uniform word line thickness, a high resistance can occur in the middle of a word line. This is much as described in US patent publication number 2012-0008405.

For either sort a break, a relatively high resistance will result along the word line so that when a voltage is applied to the word line, the portion of the word line on the far side break will charge up more slowly. For the sort of break illustrated in FIG. **36**, this will affect all of the memory cells along the word line. For the sort of break illustrated in FIG. **37**, memory cells on the far of break (to the right, as drawn in FIG. **37**) will be affected. When a voltage is applied along word line, whether a sensing or programming voltage, the control gates of the affected cells will charge up more slowly than for the non-affected cells, as illustrated in FIG. **38**.

FIG. **38** shows a programming waveform of an increasing amplitude staircase of programming pulses alternating with verify operations. The broken line illustrates the desired waveform. On the far side of a break, however, the voltage level seen will be as shown in the solid line due to the resistance at the break. The lower of the effective programming pulse, so that memory cells seeing this reduced pulse will take more pulses to write, if they can even be success-

fully programmed to verify at their target states. As the break will also affect the memory cell's gate voltages for the verify operation, so that the determination of threshold voltages after a pulse will be inaccurate. As the program pulse voltage (VPGM) and verify voltage (VFY) are both lower for the affected cells, there threshold distributions (Vt) are shifter lower, as illustrated in FIG. **39** for a four state per cell (Er, A, B, C) embodiment. In FIG. **39**, the solid lines shows the desired distribution and broken line represents the leftward shifted cells that can result from broken word lines. As both pulse and verify level are lowered, an effected cell may appear to pass (successfully verify) after some extra pulses, but still not be sufficiently programmed. Consequently, a broken word line test of this sort based on a standard program operation may not be sufficient to catch very resistive broken word lines based on the relative number of additional programming pulses as this can also occur due to process variation, with resultant yield loss and erroneous detection of broken word lines.

This section looks at two methodologies for detecting word line breakage. In a first of these, an algorithm uses a modified programming process where the verify's sensing operation is elongated. This can be used to detect broken word lines of the type associated with both FIG. **36** and FIG. **37**. This methodology can also be applied to memory structures (such as those of BiCS type) that have programmable select gates in order to determine broken select lines. In an exemplary embodiment, an internal circuit on the memory device can detect the delta of program pulses between good control lines (i.e., word lines or select gate lines) and broken control lines to identify the broken control lines. In a second set of methodologies that is useful for the sort of break illustrated with respect to FIG. **36**, differences in ramp-up times for a word line (or select line) when a VPGM or other voltage is applied can be used to differentiate good word lines (or select lines) from those with hookup related failures.

Looking at test program operation with the elongated verify, it is useful to have a good word line baseline calibration. This can be done by detecting the average number of pulses (NP) for good word lines using a standard programming staircase for the VPGM pulse values and pulse widths and standard very VFY voltages and timings. The number of pulses for a given word line can be represented as NP=N+p, where N is the average number of pulses for a good word lines and p is a variation from the average. The average N can be calibrated and record the value in internal latch. Depending on the embodiment, the N value can be the average of a block or set of blocks (such as a plane or die), where outliers can be ignored, or word line to word line variations within a block can be used, as discussed in more detail in the references cited above. If the delta p is in the variation range of, say, 2 or 3 loops from the average, then this can be considered to be process variations for normal word lines.

After the good word line calibration, broken word lines can then be detected by a test program operation to determine the number of pulses when using the target program level and pulse width, and where the verify voltage VFY is again the target value, but with an elongated verify time. A parameter can be used to change this verify time.

For reference, FIG. **40** shows the cell distribution for a properly programmed cell population in a four state per cell embodiment, where vfyA/B/C are the target verify levels. FIG. **41** is a standard program pulse/verify word line waveform. For simplicity, only single state verify is shown between pulses, although multi-state memories typically include multiple verifies corresponding to different states between at least some pulses. FIGS. **42** and **43** are the corresponding figures for the elongated verify test program operation.

As in FIG. **38**, FIG. **43** shows the applied word line waveform as the broken line, inside of which is the voltage as seen on the far side of the break. By increasing the verify time relative to a standard write, the broken word line can develop to reach the target Vfy level, so that there is no longer the shift in the distribution, although this can require a significantly higher number of VPGM pulses. In this embodiment, the broken word line verify levels (vfyAbwl, vfyBbwl, vfyCbwl) are taken the same as the corresponding standard verify levels because of the slow program verify times. The number of programming pulses for a broken word line is then NP=N+X, where is a delta above the average. The NP value can then be recorded to check for a bad word line. If the delta X is greater than, say, 3 loops, this can be reported out as a suspected broken word line failure for the word line. (If N+X=Y, for some limit Y, then a fail status would result since program loop reaches a maximum value.)

In terms of system side implementation, the memory system can monitor an in-build Delta X loop count for any abnormal word lines from the memory chips and keep a backup copy of data on those word lines only. This can then be used broken word line failures during reads that lead to an uncorrectable ECC error. This sort of arrangement can be complementary to the various post-write read and enhance post-write read (EPWR) techniques descried in US patent publication and patent numbers US-2011-0096601; 2013-0031431; U.S. Pat. Nos. 8,566,671; and 8,726,104.

FIG. **44** is a schematic representation of some of the elements on the memory chip, similar to (and similarly number to) the similar figures of the preceding sections, but with the elements pertinent to the present discussion explicitly included at **4430**. The reference value for the number of pulses expected for programming of a good word line is latched at **4431**. To determine whether a word line is bad, the number pulses used to program a word line in the test program can be latched at **4433**, with the reference value subtracted at **4435** and compared by logic at **4437**. (Although shown to be separate for the discussion of this section, these function can be part of the general control logic of the chip in block **4413**.) If the WL meets the good word line criterion (here X<3) it is marked as good, if it fails it is marked as bad and the status flags marked accordingly.

By elongating the verify time, the accuracy of determining bad lines is improved relative to the standard verify timing. At the system level, the number of pulses can be monitored and used to detect any word lines end of life. As noted, this technique can be applied to broken select gate lines as well as word lines, although a separate calibration may be used for the select gate lines.

A second of methodologies for determining broken word line looks at the rate at which the interconnection circuitry the word line decoder circuitry and the word line itself charges up when driven through the decoder. If the word line is broken at some point along this interconnection line, the high resistance at the break will restrict the current flow causing the interconnect line to charge up more quickly. Consequently, this approach can be particularly effective for breaks of the sort illustrated with respect to FIG. **36** or other breaks between the driver and the control gate portion of the word line, although it can also be used to determine breaks on the control portion near the interconnect. This is effect is illustrate in FIG. **45**.

In FIG. **45** the line **4501** represents the ramp up rate for a normal word line and **4503** represents the ramp up rate for a word line with a defect such as the under etch shown in FIG. **36**. In this example, the applied voltage is a programming voltage VPGM, since this effect will typically be more pronounced for higher voltages, but other voltage levels can be used. Note that this technique can also be applied to determining similar break in select gate lines. In any of these cases, as the broken word line ins highly resistive at or near the beginning or the word line, the driver will not see the complete capacitance resulting in the quicker ramp up. An analog circuit can then detect the time difference At between a bad word line and a good word line to reach, say, half of the applied voltage. FIG. **46** shows some of the same elements as in FIG. **21**B to show at **4601** an example of a point after the pass transistor where this slope can be measured for a word line or select gate line.

Concurrently Selection of Multiple Word Lines for High Voltage Stress

A number of stress modes can involve the application of high voltages to word lines, such as various stress modes described above or in US patent publications: 2012-0008405; 2012-0008384; 2012-0008410; 2012-0281479; 2013-0031429; 2013-0031429; and 2013-0229868. For example, control gate to substrate or floating gate to control gate shorts can be checked by programming the word lines to the highest states. Typical decoding/driving circuitry can only apply a programming or other high voltage to a single selected word line per block. For a stress mode using such a high word line voltage, having to do a single word line per block can be have a large test time impact. Alternately, concurrently selecting all word lines across multiple blocks to receive a program voltage would greatly speed up the process, but would place great demand on the charge pump circuitry providing the voltage and generally stress the memory circuit. To improve upon this situation, the present section presents a memory circuit structure and corresponding techniques with a stress mode where multiple words within the same block can concurrently be selected to receive a programming or other high voltage, while non-selected word lines can receive the pass voltage as used for non-selected word lines in a standard program operation and non-selected blocks can be left to float.

More specifically, a multi-word line select option for a given block can be used for a group of selected word lines to be set to the a programming voltage VPGM or other high voltage, while the unselected word lines of the block are set to Vpass to minimize electric field differences in order to avoid disturb. For example, a group of selected word lines could number 4, 8 or 16. In an exemplary embodiment, the multi-word line option can be applied to one block per plane, so that if there are two memory planes, for example, two such blocks can be selected simultaneously for the multi-word line option for those blocks. FIGS. **47** and **48** respectively illustrate 2D NAND and BiCS based examples.

FIG. **47** shows a representative NAND string **4701** supplied along bit line BL on the drain side through bit line select switch BLS **4703** and on the drain side to the common source line CELSRC. In this example, memory cells **0-85** are connected in series along word lines WL**0-85** between source and drain select gates SGS and SGD, where a dummy word line WLDS and WLDD is included on either end. Operating voltages for the word lines are supplied by the various bias and driving circuitry, here represented by charge pumps supplying the programming voltage **4711** and pass voltage **4713**. The decoding circuitry to selectively apply these voltages along the word lines is represented by the transfer gates connected along the line TransferG **4705**. A pair of gates **4707** and **4709** is also included as an additional path to set select gate levels. As shown, an edge group of word lines on the drain side is selected to receive Vpgm, the drain select gate is set VSGD, and unselected word lines are set to Vpass; that is, the string can be biased as typical for programming, but with more than a single word line receiving Vpgm. Here the selected word lines are an adjacent set, but other embodiments can have varying levels of interleaving between the selected and unselected word lines. Any dummy word lines can be treated as for other non-selected word lines.

FIG. **48** is a counterpart of FIG. **47** in the contest of a BiCS type memory such as described above with respect to FIGS. **9-12**. The multiple select gates SGD and SGS on either end can be biased as in a standard write. For the word lines, a set of, say, 4, 8, or 16 of these can be selected to receive concurrently the high voltage, while the rest get the pass voltage. The selected groups of word limes, where the different numbers (e.g., 4, 8, or 16) can be set by a parameter, can then be biased to Vpgm as supplied by the corresponding charge pump. The unselected word lines are biased to Vpass by the Vpass pump, or, if the circuit is using a high voltage (such as 12V) poser supply, by a drop down regulator circuit.

For any of these arrangements, being able to simultaneously stress at a high voltage more than a single word line per block can improve the defect detection process both in the number of stress modes available and in terms of the level of parallelism that can be used. In terms of the degree of parallelism obtainable from a multi-word line select of a single block, a single word lines from a similar number of blocks could be concurrently stressed and obtain the same amount of parallelism; however, not just the selected blocks are biased in this process, but also any non-selected word lines of selected blocks are also biased to the Vpass. Consequently, a multi-block, single word line would use much more current and power from the charge pumps, possibly to the point that the pump will collapse, that the equivalent single block, multi-word approach and also word lines with the Vpass level many more times.

FIG. **49** is a block diagram that highlights some circuitry involved and is similar to, and similarly numbered as, FIG. **44** and other such figures above. In particular, the control gate decoding circuitry and passed gates used to biased the word lines of the array **4901**, which can be of the 2D NAND or BiCS variety, for example, is singled out. The charge pump circuits **4923** supply the control gate decoding block **4941**. Based on the decoding signals from block **4917**, the selected word lines are the supplied with Vpgm through the high voltage switches along **4943** and transfer gates along **4945**. Vpass is similarly supplied to the non-selected word lines.

The use of this multi-word line select more can provide for better control of voltage levels when used in test modes with creating big electric fields between the selected and unselected word lines, helping to lower the amount of disturb. As multiple word lines are concurrently set high, control gate to substrate stress test modes can be greatly accelerated relative to a single word line stress, as can various word line to word line, flash write, and control gate to floating gate stress modes.

Multi-word line select can also be used to monitor critical dimension (CD) differences of word lines. Memory devices can have significant word line CD variation between early (toward source side), middle, and end (toward drain side)

word line groups. The ability to apply high program voltages concurrently to multiple word lines can help to bring this issue out at test time.

The word line grouping methodology can be used for early word line screen to detect reliability issues and also simplify word line tests at the system level as it can be applied to groups of word lines within a block. As an example of a test implementation that can be performed at die sort for word line dependency issues, some sample block can be selected in which word line groups can be programming with an elevated programming voltage. By using groups of word lines, this only needs to be done a few times: for example, if groups of 16 are used on blocks of 48 word lines, only 3 programming are needed. The pulses can then be examined to whether some word line groups are faster or slower than others.

CONCLUSION

The foregoing detailed description has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the above to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. The described embodiments were chosen in order to explain the principles involved and its practical application, to thereby enable others to best utilize the various embodiments and with various modifications as are suited to the particular use contemplated. It is intended that the scope be defined by the claims appended hereto.

It is claimed:

1. A non-volatile memory circuit, comprising:

an array of non-volatile memory cells fainted according to a NAND architecture as a plurality of blocks, each block formed of a plurality of NAND strings having multiple memory cells connected in series and connected along word lines;

bias circuitry providing bias voltage levels for use in the operation of the array; and

decoding circuitry whereby the bias circuit is connectable to the array to selectively apply the bias voltage levels thereto,

wherein, when performing a programming operation on a selected word line, the decoding circuitry applies a programming voltage to the selected word line while applying a pass voltage to the other word lines of the block to which the selected word line belongs, the programming voltage being higher that the pass voltage, and

wherein, when performing a multi-word line stress operation on a first plurality of selected word lines of a first block in which the first plurality of selected word lines are a contiguous group of word lines of the first block, the decoding circuitry applies a stress voltage concurrently to the first plurality of selected word lines while applying the pass voltage to non-selected word lines of the first block, the stress voltage being higher than the pass voltage.

2. The non-volatile memory circuit of claim **1**, wherein the bias circuitry includes a charge pump that generates the programming voltage.

3. The non-volatile memory circuit of claim **1**, wherein the bias circuitry includes a charge pump that generates the stress voltage.

4. The non-volatile memory circuit of claim **1**, wherein the bias circuitry includes a charge pump that generates the pass voltage.

5. The non-volatile memory circuit of claim **1**, wherein the stress voltage is the programming voltage.

6. The non-volatile memory circuit of claim **1**, wherein the stress voltage is higher than the programming voltage.

7. The non-volatile memory circuit of claim **1**, wherein the non-selected word lines of the first block to which the pass voltage is applied for the multi-word line stress operation are all of the non-selected word lines of the first block.

8. The non-volatile memory circuit of claim **1**, wherein the first plurality of selected word lines includes an edge word line of the selected block.

9. The non-volatile memory circuit of claim **1**, wherein the first plurality of selected word lines does not include an edge word line of the selected block.

10. The non-volatile memory circuit of claim **1**, wherein the non-volatile memory circuit performs the stress operation as part of a built in self-test operation.

11. The non-volatile memory circuit of claim **1**, wherein the non-volatile memory circuit performs the stress operation in response to an external command.

12. The non-volatile memory circuit of claim **1**, wherein the non-volatile memory circuit is a monolithic three-dimensional semiconductor memory device where the memory cells are arranged in multiple physical levels above a silicon substrate and comprise a charge storage medium.

13. The non-volatile memory circuit of claim **12**, wherein the NAND strings run in a vertical direction relative to the substrate, and the word lines run in a horizontal direction relative to the substrate.

14. The non-volatile memory circuit of claim **1**, wherein the non-volatile memory circuit is a monolithic two-dimensional semiconductor memory device where the memory cells are arranged in a single physical level.

15. The non-volatile memory circuit of claim **1**, wherein the first plurality of selected word lines of the first block is all of the word lines of the first block.

* * * * *